SENTIMENT ANALYSIS:

EMOTION, METAPHOR, ONTOLOGY & TERMINOLOGY (EMOT-08)

A workshop presentation at the LREC Conference, Marrakech

May 27th, 2008

Room 2, Palais des Congrès Mansour Eddahbi

Boulevard Mohamed VI

40 000 Marrakech

Morocco

Khurshid Ahmad (Editor)
Trinity College, Dublin

Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology

27th May 2008

WORKSHOP ORGANISERS

Khurshid Ahmad, Trinity College, Ireland (Chair)[1]
Gerhard Budin, Universitat Wien, Austria
Ann Devitt, Trinity College, Ireland
Sam Glucksberg, Princeton University, USA
Gerhard Heyer, Universitat Leipzig, Germany
Maria Teresa Musacchio, Universita di Padova, Italy
Maria Teresa Pazienza, University of Roma Tor Vergata
Margaret Rogers, University of Surrey, U.K.
Carl Vogel, Trinity College, Ireland
Yorick Wilks, University of Sheffield, U.K.

WORKSHOP PROGRAMME & PROCEEDINGS

[1] This workshop has been partially supported by the Long Room Hub Project of the Trinity College, Dublin, Ireland.

EMOT 2008 WORKSHOP PROGRAMME

09:00-09:10 Welcome and Introduction to the Workshop
Khurshid Ahmad
09:10-10:00 THEORY - Beyond Similarity: How metaphors create new categories
Sam Glucksberg

10:00-10:30 REFRESHMENT BREAK

10:30-11:00 THEORY - Metaphor as resource for conceptualisation and expression of emotion
Andrew Goatly
11:00-11:20 METHOD - Sentiment analysis using automatically labelled financial news
Michel Généreux, Thierry Poibeau and Moshe Koppel
11:20-11:40 METHOD - An unsupervised method for extracting topic expressions from reviews on TV shows
Takeshi S. Kobayakawa, Jin-Dong Kim and Jun'ichi Tsujii
11:40-12:00 APPLICATION - Automating opinion analysis in film reviews: Statistical and versus linguistic approach
Damien Poirier, Cécile Bothorel, Émilie Guimier De Neef and Marc Boullé
12:20-12:40 APPLICATION - Detecting Uncertainty in Spoken Dialogues: The use of prosodic markers
Jeroen Dral, Dirk Heylen and Rieks op den Akker

12:40-17:30 POSTER SESSION

METHOD - Revisiting the use of lexically—based features for sentiment detection
Ben Allison
APPLICATION - Annotating opinion—evaluation of blogs
Estelle Dubreil
MULTILINGUAL METAPHORS - Politics makes the Swedish :-) and the Italians
Jerom F. Janssen and Carl Vogel
APPLICATION - Motion and emotion or how to align emotional cues with game actions
Gaëlle Lortal and Catherine Mathon

13:00-14:00 LUNCH

14:00-14:30 THEORY -The deep lexical semantics of emotions
Jerry R. Hobbs and Andrew Gordon
14:30-14:50 MULTILINGUAL METAPHORS -Universal or culture-specific metaphors in economics? A corpus study of original vs translated Italian
Maria Teresa Musacchio
14:50-15:10 APPLICATION - Flame, risky discussions, no flames recognition in forums
Maria Teresa Pazienza, Armando Stellato and Alexandra Tudorache
15:10-15:30 APPLICATION - Co-word analysis for assessing consumer associations: A case study in market research
Thorsten Teichert, Gerhard Heyer, Katja Schöntag and Patrick Marif
15:30-15:50 APPLICATION - Affect transfer by metaphor for an intelligent conversational agent
Alan Wallington, Rodrigo Agerri, John Barnden, Mark Lee and Tim Rumbell

16:00-16:30 REFRESHMENTS (AND POSTER SESSION)

16:30-16:50 THEORY - Metaphor is generic
Carl Vogel
16:50-17:10 MULTILINGUAL METAPHORS-Now you see them, now you don't: Lexicalised metaphors in translation
Margaret Rogers
17:10-17:30 APPLICATION -The 'return' and 'volatility' of sentiments: An attempt to quantify the behaviour of the markets?
Khurshid Ahmad

17:30-18:00 DISCUSSION & WORKSHOP CLOSURE

Table of Contents

Introduction
Khurshid Ahmad
..……1

I. THEORY
Beyond Similarity: How metaphors create new categories
Sam Glucksberg
..……6
Metaphor as resource for conceptualisation and expression of emotion
Andrew Goatly
..……7
The deep lexical semantics of emotions
Jerry R. Hobbs and Andrew Gordon
..……16
Metaphor is generic
Carl Vogel
..……21

II. METHODS
Revisiting the use of lexically—based features for sentiment detection (Poster)
Ben Allison
..……30
Sentiment analysis using automatically labelled financial news
Michel Généreux, Thierry Poibeau and Moshe Koppel
..……38
An unsupervised method to extract topic expressions from reviews on TV shows
Takeshi S. Kobayakawa, Jin-Dong Kim and Jun'ichi Tsujii
..……44

III. MULTILINGUAL METAHPORS
Politics makes the Swedish :-) and the Italians :-((Poster)
Jerom F. Janssen and Carl Vogel
..……53
Universal or culture-specific metaphors in economics? A corpus study of original vs translated Italian
Maria Teresa Musacchio
..……62
Now you see them, now you don't: Lexicalised metaphors in translation
Margaret Rogers
..……70

IV. APPLICATIONS
Detecting Uncertainty in Spoken Dialogues: An explorative research to the automatic detection of a speakers' uncertainty by using prosodic markers
Jeroen Dral, Dirk Heylen and Rieks op den Akker
..……72
Motion and emotion or how to align emotional cues with game actions (Poster)
Gaëlle Lortal and Catherine Mathon

……79
Flame, risky discussions, no flames recognition in forums
Maria Teresa Pazienza, Armando Stellato and Alexandra Tudorache
……86
Automating opinion analysis in film reviews: The case of statistic versus linguistic approach
Damien Poirier, Cécile Bothorel, Émilie Guimier De Neef and Marc Boullé
……94
Co-word analysis for assessing consumer associations: A case study in market research
Thorsten Teichert, Gerhard Heyer, Katja Schöntag and Patrick Marif
……102
Affect transfer by metaphor for an intelligent conversational agent
Alan Wallington, Rodrigo Agerri, John Barnden, Mark Lee and Tim Rumbell
……107
The 'return' and 'volatility' of sentiments: An attempt to quantify the behaviour of the markets?
Khurshid Ahmad
……114
Annotating opinion—evaluation of blogs (Poster)
Estelle Dubreil
……124

COVER PAGE DESIGN: The background of the Cover Page comprises Chinese equivalents of the word *sentiment* (观点-pronounced: *guandian*) and *emotion* (情感-pronounced: *qinggan*). With grateful acknowledgement to Dr. Chaoxin Zheng.

Introduction: The role of emotion, metaphor, ontology, and terminology (EMOT) in *sentiment analysis*

Khurshid Ahmad,
Department of Computer Science, Trinity College, (University of Dublin),
Dublin 2, IRELAND
kahmad@cs.tcd.ie

Abstract

This workshop brings together colleagues from psychology, computer science, linguistics, including applied, computational and corpus linguistics, and translation studies, to discuss a multi-faceted topic: how ***sentiment*** is articulated in written and spoken language and is articulated across languages. Our goal in this workshop is to understand how to build and test computer systems that can analyse sentiments. This workshop covers theory, method, applications and multi-lingual aspects of sentiment analysis, with reference to topics on emotion, metaphor, terminology and ontology.

This workshop deals with the recent advances in the processing of "sentiment" in arbitrary collections of texts. Sentiment has been defined as:

> **'A mental feeling, an emotion. Now chiefly applied, and by psychologists sometimes restricted, to those feelings which involve an intellectual element or are concerned with ideal objects'. (Oxford English Dictionary)**

Sentiment can be expressed about works of art and literature, about the state of financial markets, about liking and disliking individuals, organisations, ideologies, and consumer goods. The representation (and consequently the recognition) of sentiment or emotion in text remains a matter of debate. In cognitive psychology, emotion is often defined in terms of a set of discrete (possibly universal) states or more commonly in terms of orthogonal dimensions such as evaluation and activation. The expression of sentiments in language is often characterised by 'two opposite or contradictory tendencies, opinions, or aspects' (OED), and hence some authors refer to sentiment analysis as polarity analysis. It is necessary to examine what aspects of emotional experience sentiment analysis aims to capture and how best to represent this.

In psychology and in (computational) linguistics, the notions of emotion and metaphor interact in a number of complex ways. It has been argued that conceptual metaphors underlie human understanding and processing of emotion. For example, in the debate over the environment, the choice of the term 'global warming' over the neutral 'climate change' gives (negative) emotional overtones to the issue. In addition, it can be argued that the expression of sentiments and its interpretation can rely critically on how a speaker or writer uses metaphor. An understanding of how emotion is

expressed and perceived in language is not complete without addressing the role of figurative language and metaphor as basic scaffolding or tool for modulating affective text content.

These theoretical questions address emotion in language in a general sense. Currently, sentiment analysis typically deals with a specific domain of 'ideal objects'. In order to build a sentiment analysis system, one has to understand 'what there is' in a given domain, i.e. the ontology of the domain. Once a decision has been made on the ontological basis then consideration has to be given as to how this ontology is articulated: what resources are needed to express the terminology of the domain and how to access these terms. In this context, is it possible to conceive of generic sentiment analysis? Practitioners in this area need to examine the requirements of an approach that could cross boundaries of domain or time or even language where different communities of use, languages or cultures may express or even experience sentiments in different ways. In English a measure of a thriving economy translates (metaphorically) in an invitation to 'count the construction cranes' one can see in major towns and cities, whereas in Italian reference is made to how well 'il mattone' (the brick) – as a metonymy for the building industry – is performing.

Work in sentiment analysis may be regarded as work in intelligent information retrieval and "success" is evaluated in terms of accuracy in identifying the affective content of information segments. Yet sentiment analysis has the potential to have a powerful impact in other domains that require input about their emotional context. For example, some approaches to Human Computer Interaction and emerging research in Affective Computing rely on machines having some level of understanding of emotions and sentiments. Currently, such input is dependent on the intuition of the developers of affective or expressive computing systems. The developments in sentiment analysis will help affective computing in that there will be less reliance on the intuition of the developer of the affective system. Corpus-based methods, when used together with more intuitive work in sentiment analysis, will perhaps increase the quality of the output, much in the same way as corpus-based lexicography can be used to substantiate or negate the intuition of a lexicographer. Workers in Human-Computer Interaction, Affective Computing, Lexicography and Terminography, may become end-users of work in sentiment analysis and sentiment analysis folks may have much to learn from how a machine endowed with emotions/sentiments behaves. It may become feasible to evaluate sentiment analysis systems in terms of the performance of such applications. An examination of alternative end-user systems and evaluation mechanisms can only serve to enrich the field and present new challenges for researchers to address.

This interdisciplinary workshop will address three related topics:

(a) how metaphor and sentiment interact in everyday communication;
(b) language/conceptual resources properties to support sentiment analysis
(c) evaluation of sentiment analysis programs and evaluation methodologies.

The proceedings of the EMOT-2008 workshop are divided into four overlapping parts: First there are four papers on **theory** –Sam Glucksberg's keynote contribution on the creation of new categories and metaphors is followed by Andrew Goatly's paper on the role of metaphors in the conceptualisation and expression of emotions; Jerry Hobbs and Andrew Gordon's esplore the (linguistic) semantic basis of metaphors; and, Carl Vogel's paper queries established approaches to metaphor formation and expression. The second part of the Proceedings includes work on **methods** for extracting metaphors (semi-) automatically: Moshe Koppel and colleagues have extended their work on the use of learning methods in the identification of metaphors; Takeshi Kobayakawa and colleagues explore unsupervised methods in this context; the method section includes a poster by Ben Allison's work on lexical features of words used as metaphors. The third section deals with the identification and use of metaphors in a **multilingual context**: Teresa Musacchio has looked at metaphors used in economics in English and Italian and asks whether metaphors are universal or culture specific. Margaret Rogers is looking at lexicalised metaphors and translation. Finally, the fourth section on **applications** deals with the analysis of metaphors: in blogs and e-mails (Estelle Dubreil's poster and Maria Tersea Pazienza et al's paper); in politics (see the poster presentation by Jerom Janssen and Carl Vogel); in marketing (Gerd Heyer and colleagues); film reviews and entertainment (Damien Poirier and colleagues, and a poster by Gaelle Lortall and Catherine Mathon); an innovative approach to affect transfer in conversational situations is discussed by colleagues at the University of Birmingham (Alan Wallington, John Barnden and colleagues), and we have a contribution on detecting 'uncertainty in spoken dialogues (Rieks op den Akker and colleagues). The last article is mine (Khurshid Ahmad) on quantifying affect for computing risks based on an automatic analysis of the text of financial and political news.

Sentiment analysis will play a key role in intelligent information extraction from text corpora ranging from telephone transcripts recorded surreptitiously to opinions expressed openly about consumer goods, books and films for example. In between, we have sentiments expressed covertly or overtly in order to discover the price of financial instruments, to influence a voting public or numerous other contexts. Sentiment analysis may yet turn out to be as big a challenge as machine translation, corpus-based lexicography, and information extraction were in their time. This is a truly multi-disciplinary effort which draws on and can inform research in psychology, artificial intelligence, knowledge management, human computer interaction, affective computing and has an impact on such diverse areas as film studies, homeland security, translation studies and so on.

Acknowledgements

We have had 22 submissions: we accepted 14 submissions as papers for presentation and four for presentation as posters. The organising committee of the workshop played a key role in suggesting topics and agreeing on the final shape of the workshop. My grateful thanks, in alphabetical order of their surnames: Gerhard Budin

(Universität Wien, Vienna), Ann Devitt (Trinity College, Dublin), Sam Glucksberg (Princeton University), Gerd Heyer (Universität Leipzig), Maria Teresa Musacchio (University of Padova), Maria Teresa Pazienza (University of Roma Tor Vergata), Margaret Rogers (University of Surrey, UK), Carl Vogel (Trinity College, Dublin), and Yorick Wilks (University of Sheffield, UK). I would like to thank Tony Veal (University College, Dublin) and Mícheál Mac An Airchinnigh (Trinity College, Dublin) for agreeing to review 6 of the submitted papers at very short notice. Chaoxin Zheng formatted the proceedings and Daniel Isemann helped with proof-reading of some of the articles. My grateful thanks to all.

This workshop has been partially supported by the Long Room Hub Project of the Trinity College, Dublin, Ireland.

Theory

Beyond Similarity: How Metaphors Create New Categories

Sam Glucksberg
Department of Psychology, Princeton University
Princeton, New Jersey USA, NJ08544
e-mail: samg@Princeton.edu

Abstract (only)

Since Aristotle, many writers have treated metaphors and similes as equals: any metaphor can be paraphrased as a simile, and vice-versa. This property of metaphors is the foundation of standard comparison theories of metaphor comprehension. On this view, metaphors such as 'my job is a jail' are literally false, and so cannot be directly interpreted. Instead, such "irrational" assertions are converted to similes (i.e., my job is *like* a jail) and understood as any literal comparison would be. Comparison theories rely on three assumptions: 1. Literal interpretations have unconditional priority; 2. Metaphor interpretation is optional, triggered whenever a literal interpretation fails to make sense in context; 3. Following assumptions 1 and 2, metaphor processing is not only more difficult than literal , but involves different processing mechanisms. I argue that none of the above assumptions hold. In addition, I show that metaphors cannot always be paraphrased as similes. The different forms of a metaphor – the comparison and categorical forms – have different referents. In comparison form, the metaphor vehicle refers to the literal concept, e.g., in my lawyer is like a shark, the term "shark" refers to the literal fish. In categorical form, my lawyer is a shark, "shark" refers to an abstract (metaphorical) category of predatory creatures. This difference in reference makes it possible for a metaphor and its corresponding simile to differ (a) in interpretability and (b) in meaning. Because a metaphor cannot always be understood in terms of its corresponding simile, I conclude that comparison theories of metaphor are fundamentally flawed. Metaphors can be processed directly as categorization assertions. Furthermore, when such metaphors are novel, they create new categories that are available for public discourse.

Metaphor as Resource for the Conceptualisation and Expression of Emotion

Andrew Goatly
Lingnan University, Hong Kong
goatly@ln.edu.hk

Abstract

This paper addresses one of the workshop's themes namely that 'an understanding of how emotion is expressed and perceived in language is not complete without addressing the role of figurative language and metaphor as basic scaffolding or tool for modulating affective text content.' It falls into two overlapping halves: the metaphorical perception or construction of emotion in English; and the metaphorical expression of emotion. Data is taken from a lexical database of metaphors, and gives an overview and selected examples of metaphor themes' contribution to the construction and expression of emotion.

0. Background

Cognitive linguistic (CL) accounts of metaphor were popularised by Lakoff and Johnson (1980), and are also associated with other scholars such as Turner, Sweetser, Gibbs, Steen, Kövecses, Radden and Barcelona. They stress the ubiquity and inescapability of metaphor in thought and language, and also recognise that the metaphors we use form mental structures or schemata realised by lexical sets, known variously as conceptual metaphors, root analogies or metaphor themes. In this paper I will use the latter term. These metaphor themes involve mappings between sources (vehicles) and targets (topics, tenors) and are traditionally labelled in small caps by the formula, TARGET IS SOURCE.

In the linguistics tradition where Lakoff was nurtured, there has been a tendency to intuit metaphor themes without much lexical evidence for their importance, and so to reach doubtful conclusions about, for example, the conceptualisation of emotions (Deignan 2005: 95). To remedy this ad hoc intuitive approach, I undertook research to establish in a more principled way the important metaphor themes for English.[1] The somewhat arbitrary double criteria I used are: (1) To count as significant metaphor themes should be realised by at least 6 lexical items, found in a dictionary of contemporary English; (2) There should be at least 200 tokens of this joint set of lexical items with the relevant metaphorical meaning in the Cobuild Bank of English database. The website 'Metalude' (Metaphor At Lingnan University Department of English) is the result of these endeavours[2]. It includes an interactive database of 9000+ English metaphorical lexical items, grouped by metaphor theme, and provides the data for this paper.

Part 1 is an overview of the ways in which emotion in general and specific emotions are metaphorically conceptualised in English, with a brief aside on anger in particular. Part 2 explores how English metaphor themes and their lexis contribute to the expression of emotion/evaluation.

[1] The research was funded by the Research Grants Council Hong Kong SAR, reference LC3001/99H.

[2] http: //www.ln.edu.hk/lle/cwd03/lnproject_chi/home.html. User id: <user>. Password <edumet6>

Part 1

1.0. Conceptualisation of Emotion

The major metaphor themes for conceptualising emotion in English can be organised in four loose hierarchies. The most important grouping (Figure 1) has at the top of the hierarchy EMOTION IS SENSE IMPRESSION, with the remaining members of this group directly or indirectly dependent upon it. The important theme EMOTION IS WEATHER relates to all the sense impressions, except smell. The second, much simpler, group depends upon the joint themes EMOTION IS FLUID and EMOTION IS MOVEMENT, or the movement of fluids (Figure 2). It might be possible to see an experiential connection between these two groups: WEATHER involves MOVEMENT of FLUIDS (air and water); and EMOTION IS EXPLOSION can be linked to EXPRESSION IS OUTFLOW (of gas).

The third group uses space to indicate relationship. It too may be connected to the first group since PROXIMITY, especially in early childhood, is associated with WARMTH, relating it to AFFECTION IS WARMTH in particular and EMOTION IS TOUCH more generally (Figure 3). The fourth group concerns orientational metaphor based on the vertical axis, by which emotions in general and happiness especially are conceived as being high (Figure 4). There remain, in Metalude, a number of miscellaneous metaphor themes which do not seem to form a systematic group, beyond the fact that EMOTION is concretised as a MINERAL, and then animised as various kinds of living thing—PLANT, ANIMAL (HUMAN), a human who is a PERSON CONTROLLED, and parts of a human, BODY PART/ BODY LIQUID (Figure 5). The latter is probably the vestige of medieval medicine, the doctrine of the four humours. The remaining source, DISEASE, might link with EMOTION IS SENSE IMPRESSION in group 1.

This kind of lexicological work shows that the subgroups are often cross-linked to form larger webs and schematic interactions, a complexity that Grady has attempted to remedy with his notion of primary metaphor, though at the cost of richness of imagery and psychological force. To indicate how metaphor themes might work together to create the much touted ANGER IS HOT FLUID IN A CONTAINER (for which there is very little lexical evidence), see Figure 6.

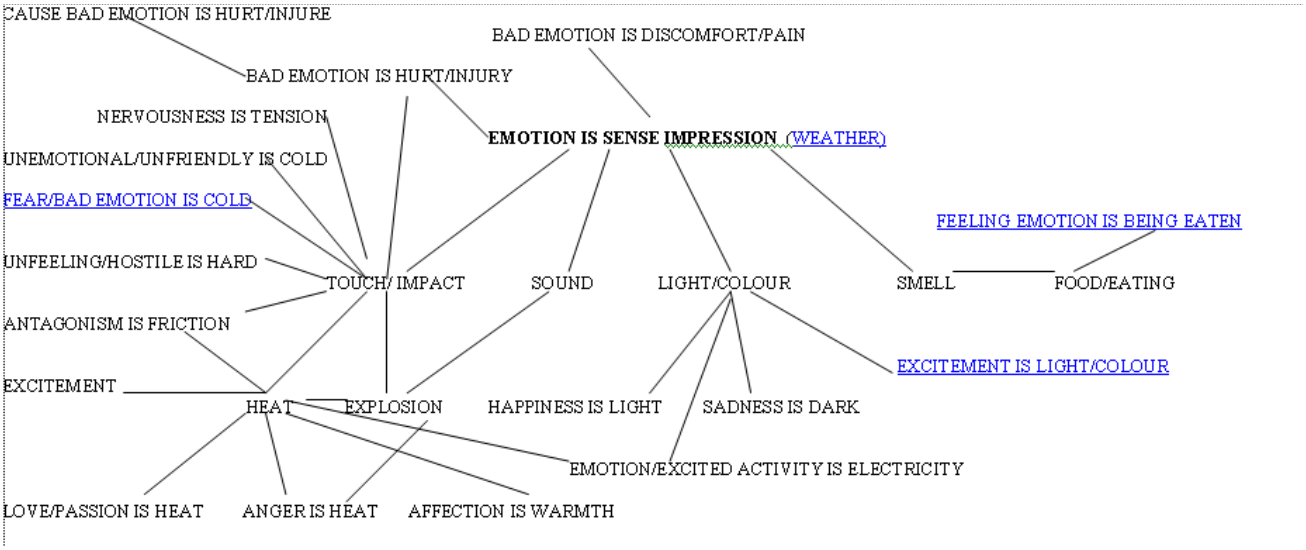

Fig 1. ***Sub-themes for EMOTION IS SENSE IMPRESSION***

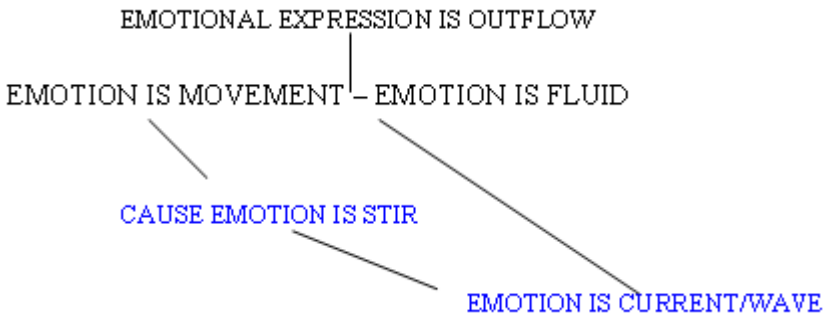

Fig 2. *EMOTION IS MOVEMENT-EMOTION IS LIQUID and their associated themes*

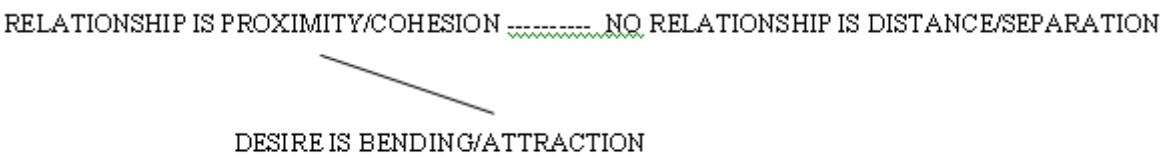

Fig 3. *RELATIONSHIP IS PROXIMITY/COHESION and its related metaphor themes*

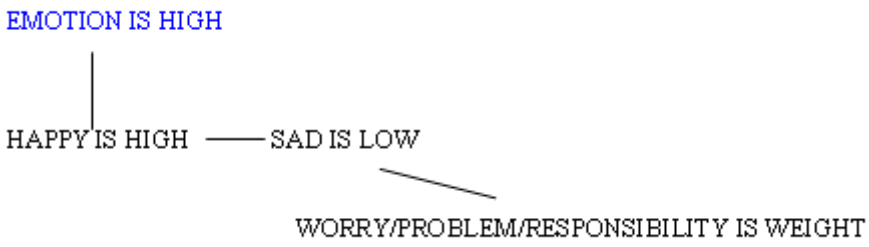

Fig 4. *EMOTION IS HIGH and its related metaphor themes*

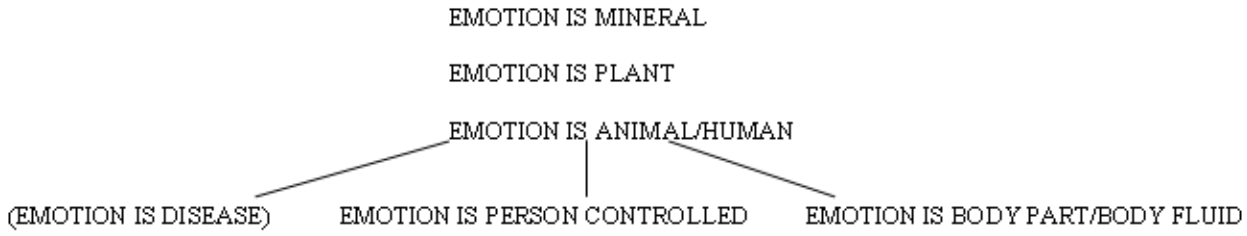

Fig 5. *Miscellaneous metaphor themes for emotions*

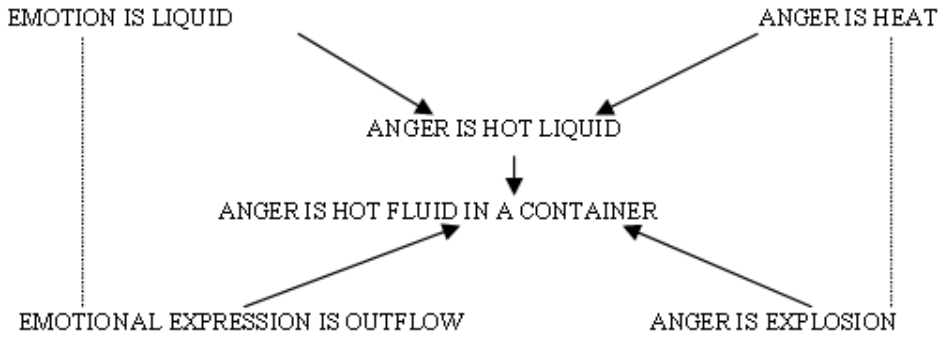

Fig 6. *Inter-relations of metaphor themes to produce ANGER IS HOT FLUID IN CONTAINER*

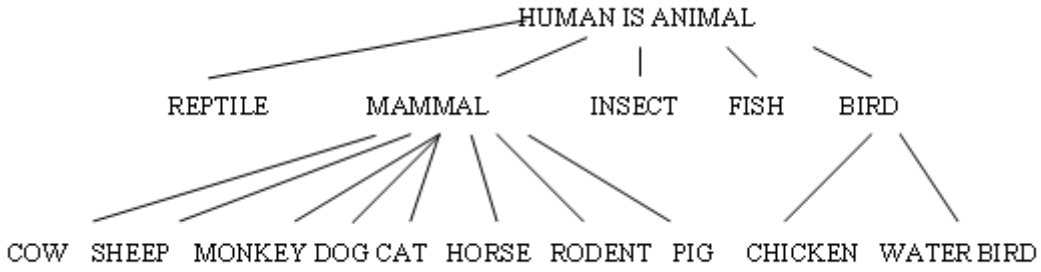

Fig 7. *Sub-themes for the metaphor theme HUMAN IS ANIMAL in Metalude*

1.1. Description and Expression of Emotion

Perhaps we should draw a distinction between the metaphorical description and metaphorical expression of emotion. As for the latter, a case can be made for regarding many swear words as metaphorical expressions of emotion. In cases like piss off, or hell, for example, the mapping or transfer of features is not a matter of conceptual or ideational meaning, but the transfer of the negative feelings about urine or eternal punishment to the meaning of the swear words. These would be clear cut cases of the expression of emotion as one of the interpersonal functions of metaphor (Goatly 1997). In the metaphorical theme EVIL/WORTHLESSNESS IS WASTE, urine and faeces, 'disgust triggers' (Eckman 2000: 174), are used as sources. Faeces metaphorise disgust and contempt for low quality: shit, turd 'contemptible, nasty person', shitty 'nasty, of low quality', shit on 'treat very badly and unkindly', pooh-pooh 'show scorn for something' (he pooh-poohed my attempts to play the piano). They are also associated with disgust for immorality: mucky 'pornographic', cesspit or cesspool 'unpleasant or immoral situation' (a cesspit of prostitution and other illegal activities). Body wastes express contempt for nonsense and uselessness: crap 'something useless, worthless, nonsensical or of bad quality' horseshit and bullshit 'nonsense', bumf (literally 'toilet paper') 'written material such as advertisements, or documents that are unwanted or boring'. The same sense of pointlessness or insignificance is found with urine: piss around 'waste time doing things without any particular purpose or plan', piddling 'insignificant' ($5 is a piddling amount).

In most of the above examples metaphorical lexis expresses rather than describes emotion, so it is interpersonal rather than conceptual, but the expression and description of emotion/evaluation often overlap. SAD IS DARK, cited previously as an example of conceptualising emotion, can describe an emotion: *the rise in interest rates* ***cast a cloud over*** ('induced pessimism about') *the property market*; *news of his father's death* ***overshadowed*** ('reduced the happiness of') *his winning the gold*. But describing, in the first person, one's own emotion, might amount to expressing it: *this is my* ***darkest hour*** ('most miserable period of my life') *and I can see no* ***light at the end of the tunnel*** ('hopes of a pleasant future situation in an unpleasant one'). Either case, expression or description/conceptualisation, involves evaluation.

Part 2.

2. 0. Metalude data for evaluation

I have sorted through the root analogies/metaphor themes listed in Metalude and attempted to extract those which seem to be evaluative. The rest of this paper is devoted to their discussion according to four questions. (1) Is there a transfer of evaluation, and if so is it from the source to the target or from the target to the source? (2) In the latter case, does this evaluative transfer depend upon the participation of the metaphor theme in some larger schema? (3) How does the selection of metaphor themes as evaluative depend upon ideological values? (4) What role does multivalency (same source for different targets) play in reinforcing evaluation?

Before I proceed, one caveat. When compiling the lexical data for Metalude, beginning back in the early 90s, my choice of metaphor theme labels was somewhat unsystematic, and heavily reliant on the traditional labels in the CL literature. Nowadays, I would be more systematic, establishing classes and hierarchies according to semantic networks or by exploiting Grady's insights into primary metaphors.

Moreover, I decided, as Metalude is conceived as a resource for teaching Chinese students English vocabulary, to subdivide metaphor themes with more than 50 lexical items. One result is that some metaphor thematic subdivisions draw attention to negative evaluations. For example, as Figure 1 shows, EMOTION IS TOUCH/IMPACT subsumes BAD EMOTION IS HURT/ INJURY. However, others, which are not sub-divided, may also contain a great deal of evaluative lexis, without this being obvious from the metaphor theme label. Clear cases are the metaphor themes with HUMAN IS ANIMAL as their superordinate (see Figure 7). Elsewhere I have shown in detail 'the negative metaphorical slant of these metaphors, many connoting unpleasantness, ugliness, pride, uncontrolled appetite and stupidity' (Goatly 2007: 152).

Metaphor themes which have no evaluative term in their title, but which nevertheless likely include evaluative terms include those in Table 1.

Table 1: ***Metaphor themes from*** **Metalude** ***incorporating evaluative lexis***

HUMAN IS ANIMAL (and its subdivisions)	HUMAN IS SUPERNATURAL/MYTHICAL BEING
EMOTION IS WEATHER	MONEY IS FOOD
EXPERIENCE IS FOOD	QUALITY IS MONEY/WEALTH
EXPERIENCE/SITUATION IS WEATHER	QUALITY IS SHAPE/SIZE
HUMAN IS SUPERNATURAL/MYTHICAL BEING	QUALITY IS TASTE/TEXTURE
KNOWLEDGE/WORDS IS FOOD AND DRINK	RANK/VALUE/CHARACTER IS METAL
LANGUAGE QUALITY IS TASTE	WEATHER IS HUMAN ACTIVITY/QUALITY

2. 1. Evaluative transfer

Some metaphor themes in Metalude have evaluations in both their source and target, for example the negative ones in Table 2.

I have already exemplified WORTHLESSNESS IS WASTE. Another, DISEASE IS WAR/INVASION, shows that some of the lexis under these themes may be positive in its evaluation, despite the pejorative nature of both sides of the label. It constructs disease pejoratively as an **attack** by **invaders** 'viruses or bacteria', or **foreign bodies** from outside. The bacteria **invade** 'enter the body', and may **strike down** 'cause illness or death to' the victims who **succumb** 'become ill'. However, the lexis also contains positive evaluations: the body may **defend** itself, **fight, combat** 'struggle to survive' the disease, through **resistance** 'immune response' and medicine might **conquer**, **vanquish** 'eliminate' a disease.

Table 2: Metaphor themes from Metalude with negative evaluation in target and source

BAD/UNIMPORTANT IS POOR/CHEAP	PROBLEM/DIFFICULTY IS DISEASE
EVIL/WORTHLESSNESS IS WASTE	AWKWARD SPEECH IS AWKWARD WALKING
BAD IS SMELLY	FAILURE IS SHIPWRECK
EVIL IS DIRT	MENTAL DISTURBANCE IS DIVISION/INCOMPLETENESS
BAD EMOTION IS DISCOMFORT/PAIN	DISEASE IS WAR/INVASION
CAUSE BAD EMOTIONS IS HURT/INJURE	

Other themes apparently transfer evaluation from the source label to the target label (table 3), a finding in tune with much of the evidence for conceptual feature mapping of individual lexical metaphors. Some of the lexis for SEX IS VIOLENCE is given in 2.3. below. As another example, consider EMOTION/IDEA IS DISEASE. Ideas and emotions can be a **bug** 'enthusiasm' that is **contagious**, **catching** or **infectious** 'easily communicated to many people'. (It is worth pointing out however that with these four lexical items the negative evaluation seems to be neutralised by the target, unlike the items below). The emotions associated with ideas can be more or less strong – **virulent** 'full of hate and fierce opposition', **pathological** 'showing extreme uncontrolled feelings'. These ideas and accompanying emotions cause harm – **poison** 'introduce a harmful idea into' the mind or are harmful **– noxious, poisonous, venomous** 'harmful, negative, unpleasant', **inflammatory** 'intentionally causing negative feelings' or **jaundiced** 'pessimistic', while negative ideas **fester** 'become more intense', like an infected wound.

Table 3: Metaphor themes from Metalude with negative transfer from source to target

FEELING EMOTION IS BEING EATEN	
EMOTION/IDEA IS DISEASE	ARGUING/CRITICISING IS WOUNDING/CUTTING
CESSATION IS DEATH	ARGUMENTS IS WEAPONS/AMMUNITION
ARGUING/CRITICISING IS ATTACKING	COMPETITION IS WAR/VIOLENCE
ARGUING/CRITICISING IS FIGHTING	SEX IS VIOLENCE
ARGUING/CRITICISING IS HITTING/PUNCHING	ARGUING/CRITICISING IS WOUNDING/CUTTING

Resistance is seen in terms of preventing disease: **sanitize** 'change in order to make it less strongly expressed or offensive', **immune** 'unable to be influenced by an idea or emotion'; or of its treatment: **cure of** 'get rid of a bad idea or emotion'.

The majority of evaluative metaphor themes have an evaluative target and an apparently neutral source, I lack the space to list them all. Here is a sample, some of which I follow up later in the paper (Table 4).

Table 4: Metaphor themes in Metalude with an evaluative target

NEGATIVE	POSITIVE
BAD IS LOW	GOOD (MORALITY, QUALITY) IS HIGH
CONFLICTING PURPOSE IS OPPOSITE DIRECTION	SHARE PURPOSE IS ALIGN
UNCERTAINTY/UNRELIABILITY IS INSTABILITY	CERTAINTY/RELIABILITY IS SOLIDITY/FIRMNESS
EVIL IS DARK/BLACK	GOOD IS CLEAN/WHITE
WORRY/PROBLEM/RESPONSIBILITY IS WEIGHT	SERIOUSNESS/IMPORTANCE IS WEIGHT
STEAL IS LIFT	UNDERSTANDING IS PENETRATION/SHARPNESS
DECEIT IS DOUBLENESS	TRUTH/CORRECTNESS IS STRAIGHTNESS

The politically-incorrect metaphor themes GOOD IS CLEAN/WHITE, and EVIL IS DARK/BLACK are clear examples, there being no intrinsic value to these two colours. Realising the first we have positive lexical items such as **white knight** 'person or organisation that rescues a company from financial difficulties', **fair** 'morally correct or just', **whiter than white** 'having a reputation

for high moral standards', **lily-white** 'faultless in character', and **whitewash** 'cover up mistakes or bad behaviour'. The second comprises mostly pejorative terms, meaning evil or wrong: black meaning 'bad' (this is a black day for the Olympics), or 'cruel or wicked' (this is a blacker crime than most I've investigated), **black and white** 'with clear distinctions between morally wrong and right', **black mark** 'fault or mistake that has been noted', **blackguard** 'a wicked person', **blackleg** 'a traitor who continues to work while other workers are on strike'; they can mean 'illegal', **black market**, **black economy**; or connote loss of reputation: black sheep 'bad person in a family who brings it into disrepute', and **blacken** 'destroy the good reputation of'.

2.2. Evaluation dependent on larger schemata

The subdivision into metaphor themes with a more manageable number of lexical items also obscures the fact that many sources are necessarily evaluated if seen as part of a larger schema, though they might not appear to be in isolation.

For instance, one of the more important superordinate schemata for conceptualising activity (life) is movement forwards, ACTIVITY/PROCESS IS MOVEMENT (FORWARD), with its more obviously positive counterparts DEVELOPING/SUCCEEDING IS MOVING FORWARD and SUCCESS/EASE IS SPEED. As sub-metaphor themes we have some in which neither target nor source are intrinsically evaluative: INACTIVITY IS IMMOBILITY; LESS ACTIVE IS SLOW; and perhaps OPPORTUNITY/ POSSIBILITY IS OPENING; PURPOSE IS DIRECTION; PURPOSELESS IS DIRECTIONLESS; SHARE PURPOSE IS ALIGN. Others quite clearly have little evaluation in their sources: CONFLICTING PURPOSE IS OPPOSITE DIRECTION—one might quite happily retrace one's steps after a walk in the country; DIFFICULTY IS MUDDY GROUND—muddy ground is ideal for planting rice; DIFFICULTY/ PREVENTION IS OBSTACLE—obstacles are an excellent barrier against threats; FAILURE/ GIVING UP IS BACKWARDS—going backwards is the preferred method of parking a car; NO DEVELOPMENT IS IMMOBILITY/ CIRCULARITY—round trip holidays always bring you back to where you started, and, indeed taking a circular route around an obstacle is an excellent example of lateral thinking as in SOLUTION IS WAY ROUND/OVER/ THROUGH; UNSUCCESSFUL /DIFFICULT IS SLOW--there is nothing intrinsically good about fast and slow (think of drinking good coffee and having good sex). But, as part of the superordinate schema, all these sources are necessarily evaluated negatively as preventing successful activity.

2.3. Ideology and evaluation

This leads us to consider our third question, already explored extensively elsewhere (Goatly 2007). Evaluation is notoriously variable and subjective, at least relative to conceptual meanings which tend to be more stable within language communities. Ideological divides within society might therefore give us different evaluative stances on the themes in Metalude. This would in some cases amount to a meta-evaluation, not just using evaluative language that represents a counter ideology but responding evaluatively to others' language use. At a deep level, for example, one might observe fundamentally opposed metaphoric models for conceiving humanity and society as in Table 5.

Table 5: Ideological and metaphorical oppositions.

Relationship	Isolation
Unity	Separation
Diversity	Sameness
Quality	Quantity
Co-operation	Competition
ORGANISATION IS MACHINE	QUALITY IS QUANTITY/SIZE
SOCIAL ORGANISATION IS BUILDING	QUALITY IS WEALTH
SOCIAL ORGANISATION IS BODY	ACTIVITY IS GAME/FIGHTING
RELATIONSHIP IS PROXIMITY/COHESION	FREEDOM IS SPACE TO MOVE

The basic distinction is between a structure, in column 1, in which the individual parts are diverse, representing different qualities and therefore incommensurate, and related to each other in a co-operative enterprise. And column 2, where the individual entities are seen as similar and therefore quantifiable, free, separate and in competition with each other. If one espouses a counter-ideology to the current late capitalist one, one might evaluate negatively uses of vocabulary which belongs to the metaphor themes in the second column.

To elaborate and exemplify further, ACTIVITY IS FIGHTING can be divided into sub-themes involving speech acts as in Figure 8. The prevalence of ARGUING /CRITICISING IS FIGHTING has provoked the following feminist and co-operative response:

> There are non-adversarial aspects of argument. And there are non-adversarial metaphors for argument – arguments may help us **build a case**, **explore a topic**, or think through a problem. Evaluating arguments may lead us to change our own minds; a critical analysis of someone else's case is not, by definition, a negative one. (Govier 1999: 7-8) [my emphasis].

Similarly ecologists (Gaia-theorists) and animal rights campaigners might give a positive evaluation to ANIMAL IS HUMAN, PLACE/LANDSCAPE IS

BODY, PLANT IS HUMAN /ANIMAL, as they seem to break down the conceptual barriers between human and non-human nature, and accord a dignity to the latter. Socialists might in principle object to the use of lexis from AFFECTION/ RELATION-SHIP IS MONEY/WEALTH and HUMAN IS VALUABLE OBJECT/COMMODITY as they wish to resist the encroachment of the market into every aspect of human life and the commodification of the human body and of relationships. Afro-Caribbeans now prefer to be known as **people of colour**, in order to replace the negative meanings of **black** (2.1) with something more exciting (EXCITEMENT IS COLOUR, Figure 1). And Feminists would certainly protest against SEX IS VIOLENCE, especially since men are usually constructed as the aggressors: **chopper, weapon, shoot his load, fire blanks, conquest, lady-killer**.

However, as with this last example, careful consideration of the specific lexis realising the metaphor theme is often necessary in order to judge its ideological affinities. Anti-materialists might be thought, for example, to resist HUMAN IS MACHINE/ IMPLEMENT. But a close look at its lexis suggests that the number of pejorative lexical items far outweigh the positive (Table 6):

Table 6: Pejorative and positive lexis in HUMAN IS MACHINE/ IMPLEMENT

PEJORATIVE METAPHORS	POSITIVE METAPHORS
crook, crock, rake, basket-case, hatchet-faced, flail, rasping; mechanically, automaton, motormouth, cog, crank, hulk	new-broom, dynamo, drive, turbocharged, high-powered, high-octane

Some of this specific pejorative lexis can be interpreted according to the anti-mechanistic dictum that reducing humans to machines or implements demeans them: we are in fact, or should be more than machines. Conservatives' and traditionalists' reactionary hackles would only be raised by careful consideration of the lexis realising UNCHANGING IS HARD/RIGID, and UNCHANGING IS STATIC:

UNCHANGING IS HARD/RIGID
NEGATIVE: **unyielding, stiff-necked, hard-line**, unwilling, stubborn or unable to change their beliefs or behaviour; **rigid, rigidity**, obstinately resisting change or persuasion; **starchy**, old-fashioned and formal in behaviour; **fossilized**, **petrified, ossified**, unable to change or develop positively; **set**, unchanging, conservative; **embedded**, unchanging, permanent; **unbending**, tending to make judgments that cannot be changed; **set/cast in stone/concrete**, extremely difficult to change
POSITIVE: **stable**, not likely to change for the worse; **solid**, certain, unwavering, loyal

UNCHANGING IS STATIC
NEGATIVE: **stuffy**, formal, boring and old-fashioned; **stick in the mud**, someone who is not willing to change or accept new ideas; **cling to**, refuse to give up a tradition or belief; **entrenched**, difficult or impossible to change **dig in/dig their heels in**, refuse to change your opinions or plans, **stuck with, tied to**, forced to accept a situation you cannot change, **stagnant, stagnate** fail/-ing to change develop or improve
POSITIVE: **settled**, permanent and predictable, **stick at, apply yourself to**, keep doing the same thing with determination despite difficulties, **stick by**, continue to give help and support to a person

In a similar way, there is nothing which alerts us to sexist ideology in the label HUMAN IS FOOD. However, the lexis indicates women are disproportionately represented as food, where their purpose is to satisfy the appetites of men: **cheesecake** 'half-naked, female, photographic models', **crackling, crumpet** 'sexually attractive woman', tart 'sexually immoral/attractive woman', **mutton dressed as lamb** 'older woman trying to look young', lollipop, peach 'attractive young girl', **arm-candy** 'attractive companion at social events'. (Though we also have **dishy, stud muffin**, and **beefcake** applied exclusively to men).

2.4. The Role of Multivalency and Opposition in Metaphor themes

As suggested elsewhere, metaphor themes can interact in interesting ways to affect our cognition and ideological value judgments, for example attitudes to immigration and race (Goatly 2007: chapter 5). What interests me in this section is the way in which multivalent metaphor themes, those with an identical source and different targets, may converge (or diverge) in terms of negative and positive evaluation. For example the four themes with straightness as source all seem positive.

GOODNESS (HONESTY) IS STRAIGHTNESS	JUSTICE/LAW IS STRAIGHT (LINE)
TRUTH/CORRECTNESS IS STRAIGHTNESS	SANITY/*NORMALITY IS STRAIGHTNESS

Table 7: ***Metaphor themes from*** **Metalude** ***involving 'height' as source.***

POSITIVE	**NEGATIVE**
HIGH	***LOW***
GOOD(QUALITY/MORALITY) IS HIGH	BAD IS LOW
HAPPY IS HIGH	SAD IS LOW
HEALTH/LIFE IS HIGH	UNHEALTHY/DEAD IS LOW
POWER/CONTROL IS ABOVE	POWERLESS/CONTROLLED IS BELOW
IMPORTANCE/STATUS IS HIGH	UNIMPORTANT/SUBORDINATE IS LOW
*MORE IS HIGH	*LESS IS LOW
(CAUSE) TO GO/BE HIGH	***(CAUSE) TO GO/BE LOW***
BE GOOD ENOUGH/BETTER IS RISE	DETERIORATE IS FALL/LOWER
GAIN POWER IS RISE	LOSE POWER/CONTROL IS DESCEND
ENCOURAGE/HELP IS SUPPORT	CONTROL IS PUSH/PUT DOWN
IMPROVE STATUS IS RAISE	REDUCE STATUS IS LOWER
ACHIEVEMENT/SUCCESS IS HIGH	FAILURE IS FALLING; FAILURE IS SINKING
*INCREASE IS RISE	*DECREASE IS FALL

Probably the most obvious pattern of significant multivalency concerns the source of height (Table 7). In cases of both STRAIGHTNESS and HEIGHT I have placed asterisks against targets which do not, at face value, seem positive or negative. Normality may be boring, having more work to do may be negative. I suggest that the sharing of sources may bring about a sharing of evaluative polarity: if GOOD IS HIGH and MORE IS HIGH, then MORE = GOOD (Goatly 2007: chapter 5). This becomes even more pronounced with the metaphor themes with multivalent source BIG.

IMPORTANT IS BIG	*FEW/LESS IS SMALL
*NUMEROUS/MORE IS BIG	*DECREASE IS CONTRACT
*INCREASE IS EXPAND	

Most of these are asterisked, and it would seem that only under the influence of IMPORTANT IS BIG do they achieve a positive evaluation for large size and negative for small size. Notice, therefore, that in none of the metaphor themes cited in this section does the source intrinsically carry an evaluation, and that evaluation is either achieved by transfer from the target, or from other targets which share the same source.

However there are some conflicting evaluations, for example WEIGHT can be given a positive evaluation as in SERIOUSNESS/ IMPORTANCE IS WEIGHT or a negative one as in WORRY/PROBLEM /RESPONSIBILITY IS WEIGHT. This difference depends upon the primary scenes or schemata and metonymic frames to which weight belongs in each case. Seriousness and importance might be associated with the weighing of goods, coins or metals, in which schema it is positive, while worry, problems or onerous responsibilities might be associated with DEVELOPMENT/ SUCCESS IS MOVEMENT FORWARDS (LIFE IS A JOURNEY), or SAD IS LOW (manifested by slouching gait, drooping shoulders downcast eyes), in which case the weight is a burden and impediment to movement or upright posture.

An important conflicting evaluation of a similar source arises with RELATIONSHIP IS PROXIMITY/COHESION and NO FREEDOM IS TYING/BINDING, one of the converses of FREEDOM IS SPACE TO MOVE (Table 5). This conflict tends to construct relationships as a loss of freedom, rather than the means of achieving identities and roles through which we are empowered, and in which 'service is perfect freedom.'(Goatly 2007).

3. Conclusion

Cognitive linguistics, as the label suggests, has for the most part concentrated on the conceptual or ideational aspects of meaning, and hence has had a great deal to say about the conceptualisation of emotion. It has, however, more or less neglected the interpersonal aspects of metaphor use, of which the expression of emotion is one. Though the lexical resources for conceptualisation/ description and expression in some cases overlap, as when 1st person description amounts to expression, in other cases, such as swear words, expression is quite distinct from conceptual meaning and depends on affective grounds. So, while the first part of this paper is treading on well-worn ground, albeit beating a lexicological rather than an intuitive path, the second part is more exploratory, and I hope, opens the way for more research. I have tried to show that data may be mined from Metalude not only to reveal how emotion is conceptualised or described but, somewhat problematically, to uncover the metaphorical evaluative resources in English.

4. References

Deignan, A. 2005. *Metaphor and Corpus Linguistics.*

Amsterdam: Benjamins.
Ekman, P. 2000. *Emotions Revealed.* London: Weidenfeld and Nicholson.
Goatly, A. 1997. *The Language of Metaphors.* London and New York: Routledge.
Goatly, A. 2007. *Washing the Brain: Metaphor and Hidden Ideology.* Amsterdam: Benjamins.
Govier, T. 1999. *The Philosophy of Argument*. Newark News.: Vale Press.
Lakoff, G. and Johnson, M. 1980. *Metaphors We Live By.* Chicago: University of Chicago Press.
Metalude. 2005. http: //www.ln.edu.hk/lle/cwd03/lnproject_chi/home.html. User id: <user>; password <edumet6>.

The Deep Lexical Semantics of Emotions

Jerry R. Hobbs, Andrew Gordon

Information Sciences Institute, Institute for Creative Technologies
University of Southern California
Marina del Rey, CA 90292, USA
hobbs@isi.edu, gordon@ict.usc.edu

Abstract

The research described here is part of a larger effort, first, to construct formal theories of a broad range of aspects of commonsense psychology, including knowledge management, the envisionment of possible courses of events, and goal-directed behavior, and, second, to link them to the English lexicon. We have identified the most common words and phrases for describing emotions in English. In this paper we describe a formalization of people's implicit theory of how emotions mediate between what they experience and what they do. We then sketch out effort to write rules that link the theory with words and phrases in the emotional lexicon.

1. Introduction

We understand discourse so well because we know so much. If we are to have natural language understanding systems that are able to deal with texts with emotional content, we must encode knowledge of human emotions for use in the systems. In particular, we must equip the system with a formal version of people's implicit theory of how emotions mediate between what they experience and what they do, and rules that link the theory with words and phrases in the emotional lexicon.

The effort we describe here is part of a larger project in knowledge-based natural language understanding to construct a collection of abstract and concrete core formal theories of fundamental phenomena, geared to language, and to define or at least characterize the most common words in English in terms of these theories (Hobbs, 2008). One collection of theories we have put a considerable amount of work into is a commonsense theory of human cognition, or how people think they think (Hobbs and Gordon, 2005). A formal theory of emotions is an important piece of this. In this paper we describe this theory and our efforts to define a number of the most common words about emotions in terms of this and other theories.

Vocabulary related to emotions has been studied extensively within the field of linguistics, with particular attention to cross-cultural differences (Athanasiadou and Tabakowska, 1998; Harkins and Wierzbicka, 2001; Wierzbicka, 1999). Within computational linguistics, there has been recent interest in creating large-scale text corpora where expressions of emotion and other private states are annotated (Wiebe et al., 2005).

In Section 2 we describe Core WordNet and our categorization of it to determine the most frequent words about cognition and emotion. In Section 3 we describe an effort to flesh out the emotional lexicon by searching a large corpus for emotional terms, so we can have some assurance of high coverage in both the core theory and the lexical items linked to it. In Section 4 we sketch the principal facets of some of the core theories. In Section 5 we describe the theory of Emotion with several examples of words characterized in terms of the theories.

2. Identifying the Core Emotion Words

WordNet (Miller, 1995; Miller et al., 2006) contains tens of thousands of synsets referring to highly specific animals, plants, chemical compounds, French mathematicians, and so on. Most of these are rarely relevant to any particular natural language understanding application. To focus on the more central words in English, the Princeton WordNet group has compiled a CoreWordNet, consisting of 4,979 synsets that express frequent and salient concepts. These were selected as follows: First, a list with the most frequent strings from the British National Corpus was automatically compiled and all WordNet synsets for these strings were pulled out. Second, two raters determined which of the senses of these strings expressed "salient" concepts (Boyd-Graber et al., 2006). CoreWordNet is downloadable from

http://wordnet.cs.princeton.edu/downloads.html.

Only nouns, verbs and adjectives were identified in that effort, but subsequently 322 adverbs were added to the list.

We classified these word senses manually into sixteen broad categories, including such classes as Composite Entities, Scales, Events, Space, Time, Communication, Microsocial (e.g., personal relationships), Macrosocial (e.g., government), Artifacts, and Economics. A very important class was Cognition, or concepts involving mental and emotional states. This included such words as *imagination, horror, rely, remind, matter, estimate*, and *idea*. Altogether 778 words senses were put into this class.

These were further divided into thirty classes based on commonsense theories of cognition we had identified from an examination of several hundred human strategies (Gordon, 2004) and had constructed formal theories of in a defeasible, first-order predicate calculus (Hobbs and Gordon, 2005). Among the thirty are theories of Knowledge Management, Memory, Goals and Plans, Envisionment (or "thinking about"), Decisions, Threat Detection, Explanations, and Emotions. 140 of the 778 cognitive word senses concern emotions, and are the focus of this paper. Some random examples of the emotion word senses are as follows (many of these are ambiguous, but it is the emotional sense that concerns us): *heart, concern, relief, anger, mood, joy, fit, embarassment, morale, apathy, pride, disgust, want,*

feel, suffer, cry, upset, provoke, terrify, fascinate, glad, exciting, happy, sympathetic, passionate, and *calmly*.

3. Filling out the Lexicon of Emotion

With the aim of providing automated tools for annotating expressions of emotion in English text, we developed a catalogue of English words and phrases that refer to emotional states and emotion-related mental events, as part of a larger effort to recognize all English expressions related to commonsense psychology (Gordon et al., 2003).

Our strategy consisted of three steps. First, we convened a group brainstorming meeting with researchers, graduate students, and administrative staff within our research lab. Participants were asked to creatively and competitively produce words and phrases that were related to emotional states, the expression of emotions, and commonsense mental processes involving emotions. The purpose of this meeting was to produce an initial list that could serve as the starting point for an exhaustive linguistic search. Second, a team of graduate students in linguistics and computational linguistics were tasked to elaborate this list by consulting a variety of thesauri, phrase dictionaries, and electronic linguistic resources. WordNet was particularly useful during this step; the list was expanded to include all hyponyms of *emotion*-1, troponyms of *provoke*-1, and troponyms of *feel*-1. Morphological derivatives of each word in the expanded list were also included, e.g., the verb *resent* relates both to its present participle (*resenting*), but also to the adjective *resentful* and its derivatives (*resentfully* and *resentment*). Third, the resulting list (several hundred emotion terms) was then organized into semantic classes by clustering terms with similar meaning. During this step, we relied heavily on the emotion categories proposed by Ortony et al. (1988), expanded by Clark Elliot (1992) to include 24 distinct emotion types. The final taxonomy added a superordinate emotion class, a class for the lack of emotion, and seven classes of terms related to emotion-related mental processes, resulting in a final list of 33 taxonomic distinctions.

In conducting this analysis, we were particularly struck by two characteristics of emotion vocabulary that distinguishes it from other terminology related to commonsense psychology, e.g. beliefs, goals and plans. First is the sheer quantity of single words that reference emotion states in the English language, in no small part due to the borrowing power of English; there are literally hundreds of words available to English-speakers to describe how they are feeling. Second is the low level of polysemy within this set; most emotion terms have only a single word sense. The list below provides several examples of each of the 33 emotion categories, with the adjectival form favored over other derivatives.

1. emotion (*affect, emotion, feeling, have feelings of*)
2. joy emotion (*blithe, cheery, comfortable, ecstatic, elated, enjoyment, happy, be in high spirits, be in Nirvana, be on cloud nine*)
3. distress emotion (*agony, bereavement, brokenhearted, cheerless, depression, despondent, sad, tearful, unhappy, be low spirited, have a sinking feeling*)
4. happy-for emotion (*glad for, pleased for, congratulatory*)
5. sorry-for emotion (*commiserative, compassionate, condolence*)
6. resentment emotion (*covetous, envious, jealous, sulky, vengeful*)
7. gloating emotion (*schadenfreude, mawkish*)
8. hope emotion (*encouragement, hopeful, optimistic, sanguine*)
9. fear emotion (*anxious, apprehensive, bode, consternation, despair, fearful, terror, timid, trepidation, uneasy, worried, have cold feet, gives one the creeps*)
10. satisfaction emotion (*consolation, delightful, gratification, pleasure, ravishment, satisfaction, solace, have a silver lining*)
11. fears confirmed emotion (*fears have come true, fears realized*)
12. relief emotion (*alleviation, assuagement, relief*)
13. disappointment emotion (*defeat, disappointment, frustration*)
14. pride emotion (*conceited, egotistic, proud, prideful, vain*)
15. self-reproach emotion (*chagrin, discomfit, embarrassment, humble, humility, meek, repentance, self-conscious, self-depreciation, shame*)
16. appreciation emotion (*appreciative, thankful*)
17. reproach emotion (*disapproval, reproachful*)
18. gratitude emotion (*grateful*)
19. anger emotion (*aggravation, angry, annoyance, belligerent, furious, pique, rage*)
20. gratification emotion (*gratifying*)
21. remorse emotion (*guilt, regretful, remorseful, rueful*)
22. liking emotion (*fancy, fascination, fondness, partiality, penchant, predilection, have a taste for, have a weakness for*)
23. disliking emotion (*abhorrent, abomination, detestable, disinclination, dislikable, execration, loathsome, repugnant, repulsive, revulsion*)
24. love emotion (*adoration, agape, amorous, devotion, enamor, infatuation, lovable*)
25. hate emotion (*animosity, bitterness, despise, hateful, malefic, malevolent, malicious, spite, venomous, have bad blood*)
26. emotional state (*mood, way one feels, how one is feeling*)

27. emotional state explanation (*reason for feeling, why one feels, cause of the emotion*)

28. emotional state change (*a shift in mood*)

29. appraisal (*assess one's emotions, figure out how one feels about*)

30. coping strategy (*way of dealing with, coping technique*)

31. coping (*dealing with the feeling, coming to terms with*)

32. emotional tendency (*emotional, moodiness, passionate, sentimentality*)

33. no emotion (*aloof, ambivalent, austere, calm, cold-hearted, emotionless, heartless, impassive, indifferent, phlegmatic*)

4. Some Core Theories

We use first-order logic for encoding axioms in our commonsense theories, in the syntax of Common Logic (Menzel et al., 2008). Since human cognition concerns itself with actual and possible events and states, which we refer to as eventualities, we reify these and treat them in the logic as ordinary individuals. Similarly, we treat sets as ordinary individuals and axiomatize naive set theory. Most axioms are only normally true, and we thus have an approach to defeasibility—proofs can be defeated by better proofs. Our approach to defeasibility is based on weighted abduction (Hobbs et al., 1993) and is similar to McCarthy's circumscription (McCarthy, 1980), but the content of the theories should survive a translation to any other adequate framework for defeasibility.

The theories of cognition rest on sixteen background theories. Included among these is a theory of scales that provides means of talking about partial orderings, the figure-ground relation of placing some external thing *at* a point on a scale, and qualitative regions identifying the *high* and *low* regions of a scale. The latter are linked to the theory of functionality mentioned below; often when we call something tall, we mean tall enough for some purpose. They also need to be linked to an as-yet undeveloped commonsense theory of distributions.

In addition, we have theories of change of state, causality, and time. The theory of causality tries to provide a defeasible notion of *cause* that can be used in lexical semantics (Hobbs, 2005). The theory of time explicates such predicates as *before*, *atTime* relating an event to a time, and a *meets* relation between intervals (Hobbs and Pan, 2004).

For this paper the most relevant cognitive theories are Knowledge Management, Goals and Planning, and Envisionment. In the theory of Knowledge Management, we characterize belief and graded belief and their relation to perception, inference, and action. Briefly, perceiving is believing, we can defeasibly do logic inside belief contexts, and our beliefs influence our actions. We also axiomatize change of belief, mutual belief, assuming, varieties of inference, justification, knowledge domains, expertise, and other similar concepts in this theory.

The theory of Goals and Planning posits agents that have a top-level goal "to thrive", have various beliefs about what will cause them to thrive and other causal knowledge, and continually plan and replan to achieve this top-level goal. Planning uses axioms about what eventualities cause or enable what other eventualities to generate subgoals of goals, and subgoals of the subgoals, until arriving at executable actions. Shared goals and plans are defined in terms of mutual knowledge and of sets of agents having goals where the shared plans bottom out in actions by individual members. We define notions of eventualities being good for or bad for an agent or group of agents relative to their goals. The function and roles of artifacts and organizations are characterized in terms of agents' goals, where the structure of the artifact or organization reflects the structure of the plan to achieve the goals. We also explicate here the notions of attempting to achieve a goal and actually achieving it. A *threat* is an eventuality that may cause one's goals not to be achieved.

The theory of Envisionment is an attempt to begin to capture what it is to think about something, particularly, in a causal manner. To envision is to entertain in one's focus of attention a sequence of causally linked sets of eventualities. For example, the Common Logic expression `(envisionFromTo a s1 s2)` says that an agent `a` envisions a sequence of causally connected situations starting with `s1` and ending with `s2`. Explanation, prediction, and planning are varieties of envisionment.

5. The Theory and Lexical Semantics of Emotion

Our theory of Emotions attempts to characterize twenty-six basic emotions in terms of the abstract situations that cause them and the abstract classes of behavior they trigger. That is, emotions are viewed primarily as mediating between perception and action. Our treatment is based in part, but only in part, on that of Ortony et al. (1988). We attempt, in addition, to axiomatize the notion of the *intensity* of emotion, and give a somewhat more central role to the "raw emotions", as described below.

Natural language is very rich in emotional terminology, and our formal theory of emotion tracks language very closely. Thus, in explicating the concepts of the theory, we are also providing the deep lexical semantics of English emotional terms. Of course, the converse is not also true; there are many more English emotional terms than would be basic predicates in an underlying theory of emotion; these others we characterize in terms of the basic predicates.

Happiness is normally caused by the belief that one's goals are being satisfied. This of course is not always the explanation of one's happiness. Imagining you will win the lottery can cheer you up, sometimes you feel happy for no identifiable reason at all, and sometimes you are unhappy even though everything is going well. This is an illustration of why virtually all the rules in the cognitive theories are defeasible.

To give a flavor of the rules in the theories, we include the fairly complex one characterizing one of the sources of happiness.

```
(forall (a g e1 e2 e3 t1 t2)
  (if (and (goal' e1 g a)
           (atTime e1 t1)
           (atTime' e2 g t2)
           (believe' e3 a e2)
           (atTime e3 t1)
           (intMeets t1 t2) <etc>)
      (exists (e4)
         (and (happy' e4 a)
              (atTime e4 t1)
              (cause e3 e4)))))
```

That is, if during time interval `t1` agent `a` has the goal `g` and believes that it will be satisfied during interval `t2`, where `t2` begins when `t1` ends, then this belief will cause `a` to be happy during interval `t1`. More succinctly, anticipating success makes us happy. The `<etc>` is an abbreviation indicating defeasibility.

An inference one can draw from one's success in satisfying one's goals is that the rules or beliefs that generate one's behavior are functional. They are the right rules. Therefore, there are two conclusions with respect to one's actions. Since the rules are correct, there will be a reluctance to change one's beliefs, at least in the relevant knowledge domains. The current beliefs are doing a good job. And one will be inclined to act on one's current beliefs. One will exhibit a greater level of activity.

Sadness is given a corresponding characterization. It is normally caused by the belief that one's goals are not being satisfied. It tends to suppress the urge to action, since one would be acting on beliefs that have shown themselves to be dysfunctional. Moreover, sadness opens one to a change in beliefs.

We have axiomatized Ortony et al.'s (1988) cognitive elaborations on basic emotions. Happiness and sorrow for someone else, resentment, and gloating are defined in terms of eventualities being good for or bad for in-groups and out-groups, where in-groups are defined in terms of shared goals. Anticipation is defined in terms of envisionment; satisfaction, "fears confirmed", disappointment, and relief are defined in terms of anticipated eventualites that are good for or bad for the agent being realized or frustrated. Pride, self-reproach, appreciation, reproach, gratification, remorse, gratitude and a certain kind of anger are defined in terms of eventualities that are good for or bad for one's self or others being merely attempted or succeeding.

Although we do not "define" emotional intensity, we do constrain its interpretations with axioms that say in some special circumstances what sort of emotions will normally be more intense than others, *ceteris paribus*. For example, normally the more salient the stimulus, the more intense the emotion, and the more intense the emotion the more extreme the response. *Intense* then labels the functionally and distributionally high region of that scale.

Our treatment of the three "raw" emotions, anger, fear, and disgust, depends on the notions of eliminating or avoiding threats. One eliminates a threat by causing a change of state (or location) in it. One avoids a threat by causing a change of state (or location) in one's self. In either case, the effect is a reduction of the threat. Anger and fear are both caused by threats. In anger, our response to it is normally to try to eliminate the threat. In fear, our response is normally to try to avoid the threat.

Fear and anger are responses to external threats. Disgust is a response to a threat that is interior, and it triggers an effort to eject the threat. "Interior" may be interpreted literally with respect to the body—most of the ways of talking about disgust involve distaste or nausea. Or we may interpret it metaphorically as referring to an in-group.

All of this is of course quite naive if viewed as a *real* theory of emotions. But we believe it is reasonable as a commonsense theory, and will allow natural language systems to make sense of most occurrences of emotion terms in English discourse.

Having explicated the basic emotions formally, we are now able to write axioms characterizing the meanings of the less central emotional terminology of English. For example, to "terrify" someone is to cause one to feel intense fear. The various emotional word senses of "calm" in WordNet can be characterized in terms of feeling or causing low emotional intensity.

There are five noun senses of *pride* in WordNet. *pride-N2* includes the Ortony et al.'s (1988) sense we characterized above as what one feels on an attempt to do something good, but also includes the feeling on success and the feeling about another person's attempt or success. *pride-N1* is a version of *pride-N2*, generalized over time. *pride-N3* refers to the causal power of *pride-N1* in one's actions. *pride-N5* is *pride-N1* carried to excess. (The fourth sense is a group of lions.) The single verb sense of *pride* means to feel or express *pride-N1*.

6. Summary

Natural language understanding requires a large knowledge base of commonsense knowledge that explicates concepts in coherent theories and links lexical items with these theories. In order to achieve high accuracy, high complexity results, this effort must be manual (as indeed dictionaries are constructed manually). Early efforts will have the most impact if done for the most central concepts and the most common word senses.

In this paper we have outlined our work in constructing background theories and theories of general cognition, and we have described in more detail the structure of the theory of Emotion, indicating how it can be used to explicate the emotional vocabulary of English.

7. References

Athanasiadou, Angeliki and Elzbieta Tabakowska (eds.), 1998. *Speaking of Emotions: Conceptualisation and Expression*. Mouton de Gruyter, Berlin, New York.

Boyd-Graber, Jordan, Christiane Fellbaum, Dan Osherson, and Robert Schapire, 2006. Adding dense, weighted, connections to WordNet. *Proceedings*, Third Global WordNet Meeting, Jeju Island, Korea, January 2006.

Elliott, Clark, 1992. *The affective reasoner: A process model of emotions in a multi-agent system*. Ph.D. Dis-

sertation, Northwestern University. The Institute for the Learning Sciences, Technical Report No. 32.

Gordon, Andrew S., 2004. *Strategy Representation: An Analysis of Planning Knowledge*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Gordon, Andrew S., Abe Kazemzadeh, Anish Nair, and Milena Petrova, 2003. Recognizing expressions of commonsense psychology in English text. *Proceedings*, 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), Sapporo, Japan, July 2003.

Harkins, Jean, and Anna Wierzbicka (Eds.), 2001. *Emotions in Crosslinguistic Perspective*. Mouton de Gruyter, Berlin, New York.

Hobbs, Jerry R., 2005. Toward a useful notion of causality for lexical semantics. *Journal of Semantics*, 22:181-209.

Hobbs, Jerry R., 2008. Deep lexical semantics. *Proceedings*, 9th International Conference on Intelligent Text Processing and Computational Linguistics, Haifa, Israel, February 2008.

Hobbs, Jerry R., and Andrew S. Gordon, 2005. Encoding knowledge of commonsense psychology. *Proceedings*, 7th International Symposium on Logical Formalizations of Commonsense Reasoning, Corfu, Greece, pp. 107-114, May 2005.

Hobbs, Jerry R. and Feng Pan, 2004. An ontology of time for the Semantic Web. *ACM Transactions on Asian Language Information Processing*, 3:66-85.

Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin, 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69-142.

McCarthy, John, 1980. Circumscription: A Form of Non-monotonic Reasoning. *Artificial Intelligence*, 13:27-39.

Menzel, Chris, et al., 2008. Common Logic Standard. http://cl.tamu.edu/.

Miller, George, 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38:39-41.

Miller, George, Christiane Fellbaum, and Randee Tengi, 2006. WordNet: A Lexical Database for the English Language. http://wordnet.princeton.edu/.

Ortony, Andrew, Gerald L. Clore, and Allan Collins, 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, New York.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie, 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165-210.

Wierzbicka, Anna, 1999. *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press, Cambridge, England.

Metaphor is Generic

Carl Vogel

Computational Linguistics Group
Intelligent Systems Laboratory
O'Reilly Institute
Trinity College Dublin
Dublin 2, Ireland
vogel@tcd.ie

Abstract

An approach to sense extension tailored for polysemy associated with non-literal language expanded to include belief revision generally. The relationship between metaphor and genericity as rhetorical devices is discussed, and both are accounted for as related species within the same framework of dynamic semantics. The theoretical apparatus is related to a dominant theory of metaphor interpretation and processing which holds metaphorical utterances to be class inclusion statements involving dual reference for the the metaphorical vehicle.

1. Background

The aim of this paper is to closely link the theory of metaphor interpretation to that of natural language generics. Both forms of expression have curious truth conditions, and it is argued that both can be understood in terms of forms of belief revision in first order languages augmented with sense distinctions. Influenced by work in dynamic semantics that formalized accounts of anaphora in discourse as eliminating possible models of sentences with pronouns, on the basis of restricting assignment functions that map variables into the domain, as pronouns are resolved to potential antecedents (Kamp and Reyle, 1993; Groenendijk and Stokhof, 1991), as well as research in belief revision (Alchourrón et al., 1985), Lemon (1998) proposed a framework for first-order logical languages which admitted both information increase and retraction ("updates" and "downdates", respectively).

Vogel (2001) proposed a comparable system for information increase only, but with the additional dimension of intensionality in that indices for interpretation were provided to account for the multiplicity of senses that a predicate name or name of individuals might have. In particular, that system provided for static interpretation which is classical, if relativized to sense, and dynamic interpretation, which in all but certain well-defined syntactic and semantic contexts has the capacity to update the characteristic functions of sets corresponding to the denotation of relation names and constants. A feature of this system is that metaphoricity is captured as a partial order that classifies indices, thus accommodating the intuition that today's novel metaphor is tomorrow's conventionalized non-literal expression, and the next day's dead metaphor, literal language. The system took advantage of the fact that natural languages supply mechanisms to indicate that non-literal interpretation is intended. For example, it has been noted that the appearance of "literally" in a sentence is a fairly reliable indicator that the sentence it appear in is not to be interpreted literally (Goatly, 1997). It also supposed that languages have internal means to support the disambiguation of the intended sense of an expression (even if the latter are periphrastic, for example, "I mean 'bank' in the sense of 'a financial institution"'). The intent was to offer a proof-of-concept response to Davidson's claim that metaphor is not within the remit of semantics, but of pragmatics (Davidson, 1984). Vogel (2001) provided a truth-functional compositional semantics that could accommodate metaphor and sense extension (expansion of predicates to new entities, and multiple senses for names of entities and relations), but rejected Davidson's claim that "special senses" are not involved in metaphoricity.[1]

In contrast, Vogel and McGillion (2002) argued that natural language generics, phenomena well studied in the formal semantics of natural language (Krifka et al., 1995; Carlson and Pelletier, 1995; Cohen, 1999; Cohen, 2001), are not in the remit of semantics but of mathematical formulation of a cognitive theory of concepts. The basis of this argument is that unlike the case of metaphor, there are no overt markers of genericity. While there is ample treatment of the ability of definite NPs, bare plurals, mass nouns and even indefinite singulars to sustain generic readings, they do not demand them. This argument essentially ignores the possibility that actually the unmarked case is generic reference, such as in determinerless classifier languages where the specific reading is what may optionally be marked as such if context of use does not clarify.

(1) Hurricanes happen in the Atlantic and Caribbean.

(2) Leslie smoked cigarettes.

(3) Leslie smoked three cigarettes.

Habituals (1) with unbounded subjects, and comparable constructions with terminative aspect (see Verkuyl (1993)) make this more clear: without a specific bound or clear definite marking on the object NP in (2), the preference is to understand the sentence as a past tense habitual, a form of generic. On the other hand, (3) exhibits terminative aspect. The test between the two potential readings is in whether

[1] Van Genabith (2001) resisted the idea of a "designer logic" and attempted an account based on higher-order logic with types and a translation of all metaphors into corresponding similes, a move disfavored by many.

the sentence tolerates modification by "for a day" or "in a day"—(2) can be continued with "for a day" but not "in a day", and (3) has the reverse pattern. To obtain the specific episodic reading, explicit marking is necessary on the object NP.[2]

The purpose of this paper is threefold. Firstly, the paper intends to improve upon the dynamic semantics provided by Vogel (2001) to account for aspects of metaphoricity, by incorporating downdates, and thus more clearly separating the dynamics of information assertion and retraction from the orthogonal dimension of metaphoricity. Secondly, the paper argues a close relationship between metaphoricity and genericity (the former is expansive, and the latter is restrictive in subsequent interpretation potential). This move, as suggested by the title, resonates with one dominant theory of metaphor understanding that holds metaphors to be class inclusion statements (Glucksberg and Keysar, 1993; Glucksberg, 2001). Thus, the third purpose is to show how the semantic approach put forward here is compatible with important aspects of Glucksberg's theory.

One important aspect of the theory is the emphasis on a difference in interpretation requirements between novel and established metaphors. The main explanatory mechanism of the theory is allowance of dual reference in the vehicle of a metaphor in its predication of the topic, ambiguous in predication of the topic between literal reference and an abstraction over that reference that retains salient attributable properties. Asymmetries of metaphors (in contrast to the symmetry of similes) are anchored in the distinction between given and new information, with respect to qualifiable dimensions in the given information and potential attributions supplied by the new information.

The next section of the paper spells out the formal system for update and downdate which is richer than the starting point provided by Lemon (1998) in a few respects (it does not require that every element in the domain have a name in the language; it admits multiplicity of sense; it admits sense designation into the language) and is conceptually more complete than the framework provided by Vogel (2001) in forcing a clearer separation between information assertion and retraction and the role of metaphoricity. Then, §3. demonstrates the relationship between the resulting system and the restricted quantification of genericity (essentially, generics are also treated as special non-literal senses). Finally, the paper shows how some of the desiderata of Glucksberg's theory are met. Others of them (for example, conflation of subject-object asymmetry in metaphors with topic-comment information packaging) are disputed.

2. Dynamics of first-order information

2.1. Intuitions about revision

To a child learning about the world via science documentaries broadcast on television, it may be news that (4) is true. The literal truth of the statement is about NPs at the same level of abstraction.

(4) A whale is a mammal.

(5) A whale is like a mammal.

Even if the sentence is provided as a voice accompanying a picture of two whales, such that the child anchors the subject NP to one of the two whales arbitrarily, (4) remains a literally true statement. As an accepted piece of news, the child extends whatever meaning of "mammal" was in place before, with the new information that one or more whales is also in that set. If the child knows that whales are not fish, the child may retract the prior creative hypothesis that the swimming fish-like thing is not a fish. Note that (5) is also true because whales are mammals, and things are generally like themselves.[3] Moreover, (5) is reversible. Glucksberg notes that metaphors are not only asymmetric, they are also sometimes only reversible with a change of meaning into a different metaphor. Glucksberg (2001, pg. 45) notes the difference between (6) and (7).

(6) Some surgeons are butchers.

(7) Some butchers are surgeons.

The former presumably has negative connotations, and the latter, positive. Later the issue of reversibility returns with emphasis on the fact that the constraint is not simply on the linear presentation of topic and vehicle (see (48)).

However, (8) is also felicitous if it is taken to mean that a specific kind of mammal is the kind "whale", and if it is taken to mean that a particular individual mammal is of the whale sort, or maybe least likely if a particular specific indefinite is both a mammal and a whale.

(8) A mammal is a whale.

Duality of reference is not unique to metaphorical expressions. Or, perhaps, generics are metaphors.

The point of the example (4) is to show that there are needs for asserting and retracting information about entities and relationships that hold among entities in the world, independently of whether the utterance accepted as effecting the change fits criteria for some figure of speech or other. A mechanism for assertion and retraction is a necessary part of information processing.

2.2. A formal model of first-order belief revision

Lemon (1998) provided a framework for modelling first-order belief revision of incomplete theories. A theory is understood in this framework as a set of agent beliefs about the world and the individuals and first-order relations within it. An agent can obtain new beliefs or retract old ones. Beliefs may be about the truth of propositions or of properties holding of named individuals. A common simplifying assumption is made that every individual in the domain has a name (Gamut, 1991). Additional beliefs may include quantificational statements, and in fact may be about any well formed sentence in a standard first order language. Beliefs, quantificational or not, may be added or subtracted. Rationality postulates are provided to ensure a consistent belief state under deductive closure.

[2] Glasbey (2007) notes that aspectual class can diverge between literal and non-literal readings of idiomatic expressions.

[3] It is felicitous for someone to say, "He is not like himself today."

> In the semantics, theory growth is modelled by way of model elimination (information update, written "$s \llbracket \phi \rrbracket$ " for state s and formula ϕ), and theory contraction ($T \dot{-} \phi$) is modelled by "downdates" (denoted "$s \rrbracket \phi \llbracket$ ") which involve *rational* model *construction*. Revision ($T \dot{+} \phi$), a consistency-preserving update, is a combination of these operations, denoted $s \llbracket \phi \rrbracket$ in the semantics (Lemon, 1998, pg. 86).

In retracting a belief from a theory, in general there will not be a unique subtheory of T that fails to entail the retracted formula (e.g. ϕ). Lemon refers to maximal subtheories of T with that status as, $T \perp \phi$, and defines a choice function α to pick out members of that set, and an intersection over all possible choices yields a total retraction of the formula ϕ from the theory T. To retract a universally quantified formula involves total retraction of a single formula in which the quantifier is removed and free instances of the erstwhile bound variable are substituted with a constant, the name of the individual which causes the universal to be retracted. Total retraction of an existentially quantified formula similarly requires retraction of all formulas obtained by substitution of each constant for now free instances of the formerly bound variable. This method works because of the substitutional approach taken to quantification. Names are taken as rigid designators and the naming of individuals in the domain is only ever monotonically increasing—it is not possible to un-name an individual, although individuals may have more than one name.

2.3. First-order belief revision adapted to sense extension

Assume a first order language, let the language have a denumerable set of constants, C, a supply of variables V, predicate names $\mathcal{R}$, indications of sense M, and the usual logical connectives. An indication of sense may be provided periphrastically and/or deixis accompanying an utterance; sense indications may supply information about reference.

(9) If c is a constant and m is an indication of sense, then c_m is a constant.

(10) If P is an n-ary predicate name, $n \geq 0$, and m is an indication of sense then P_m is a predicate name.

(11) The usual combination rules with respect to forming predications of n-tuples, complex formulae and sentences apply

As constructed, a predication (including those applied to zero arguments), may be accompanied by an indication of the sense in which it is to be interpreted, and the same for constants.

In general, dynamic semantics supposes that there is an input to interpretation and that the output of interpretation can be a truth value, but also a change in the model of the world that is input to interpretation of subsequent utterances. In classical logic, one thinks of a meaning function defined for arbitrary sentences relativized to a model which consists of a domain and interpretation function. Assuming a fixed domain, with dynamic interpretation, relativization is to the input and output interpretation function. Thus, a basic meaning function is going to be annotated with the input and output interpretation functions (as well as assignment functions for free variables), accordingly. In the case of static interpretation, the inputs and outputs are identical. In the case of dynamic interpretation, there can be an expanded or contracted interpretation function.

In extensional treatments of semantics, the interpretation of a predicate is the set of tuples each of which the predicate is true of; the interpretation of a constant is some element of the domain. Suppose the simple world of integeters given by the domain D in (12).

(12) $D = \{1, 2, 3\}$

(13) $C = \{i, ii, iii\}$

(14) $\mathcal{R}$ = {even, odd, lucky, $<$ }

A standard interpretation of the Roman numeral system might interpret the constants in (13) as (15) in functional notation, or equivalently as the set of tuples that constitute that function as in (16). The proper subset symbol (the symbol $\subset$ is used rather than $\subseteq$) makes clear that this is a proper subset of the interpretation function I: the predicates in (14) require interpretation as well, and thus supply another subset of I.

(15) $I(i) = 1$
$I(ii) = 2$
$I(iii) = 3$

(16) $\{\langle i, 1\rangle, \langle ii, 2\rangle, \langle iii, 3\rangle\} \subset I$

Similarly, the meanings of relations are spelled out in terms of the entities that stand in the relations. It is equivalent to provide them as functions or as the appropriate sets of tuples as in (17)–(20).

(17) $I(even) = \{2\}$
$\{\langle even, 2\rangle\} \subset I$

(18) $I(odd) = \{1, 3\}$
$\{\langle odd, 1\rangle, \langle odd, 3\rangle\} \subset I$

(19) $I(lucky) = \{2\}$
$\{\langle lucky, 2\rangle\} \subset I$

(20) $I(<) = \{\langle 1, 2\rangle, \langle 1, 3\rangle, \langle 2, 3\rangle\}$
$\{\langle <, 1, 2\rangle, \langle <, 1, 3\rangle, \langle <, 1, 4\rangle, \langle <, 2, 3\rangle\} \subset I$

There are no other tuples in I for any language, at the outset, besides those spelled out for the basic terms and predicates. Updating or downdating with the language means adding tuples to or subtracting tuples from the interpretation function for the language.

Additional parameters are needed for interpretation to accommodate multiple senses. So, consider the one place predicate, "lucky". One sense of this expression is in terms of chance—at some moment in time 2 might be a fortuitous outcome for some event, like a draw of a card from a deck.

Another sense is in terms of omens—seeing three seagulls on the bow might have some relevant meaning to seafarers. Thus, the meaning of "lucky" (19) might be relativized to the appropriate sense as in (21) or (22).

(21) $\{\langle lucky, chance, 2\rangle\} \subset I$

(22) $\{\langle lucky, omen, 3\rangle\} \subset I$

For the purposes of this paper, novel uses of expressions involve the creation of new senses for predicates. Sense extension involves the accumulation of new tuples in the interpretation for a predicate relative to a given sense. Belief revision in general, through dynamic interpretation, is modelled by allowing that tuples may be added or subtracted, and the logical closure computed.

In what follows, the meaning function ($[\![.]\!]$) is spelled out. The construction stipulates what arbitrary sentences of the language should mean, relativized to a model, which includes the domain and interpretation function for expressions of the language. Because the system should ultimately be dynamic in that the interpretation function is altered as expressions are analyzed, the function is annotated with the input interpretation on the left, and output interpretation on the right (${}^{I}[\![\pi]\!]^{O}$). Because the system is a first order one, assignment functions are provided for the interpretation of variables. These function like contexts that provide the reference of pronouns. Two additional aspects of context also anchor the interpretation—the default sense of an expression and the default 'world' in which interpretation is happening.[4]

2.3.1. Sense-relative static interpretation

Interpretation is relative to models consisting of a domain of entities and an interpretation function I for basic expressions in the language, which is presented in terms of the tuples comprising it. An important parameter of interpretation function is the index at which a basic expression is to be interpreted.

Let W be a collection of possible senses. Sense selection functions s map sense indicators to indices. That is, basic expressions must be interpreted within a model relative to the sense of the expression at stake, either signalled or fixed by default. Assignment functions g map variables to elements of the domain; this is the alternative to interpretation of variables via substitution of constants (Gamut, 1991). Constrain the basic interpretation function, I, as follows in (23)–(32).

(23) $\forall c \in C, w \in \mathcal{W}, \exists\, \dot{d} \in D : \langle c, w, d\rangle \in I$.

(24) $\forall P^n \in \mathcal{R}, n \geq 0, \forall \tau \in D^n, \langle P, w\rangle \oplus \tau \in I$ iff P is true of the tuple τ at index w.

The constraint in (23) indicates that for every constant and every sense, there is a unique element of the domain that the constant can denote in the provided sense; (24) provides the list of tuples of entities in the domain that constitute a relation as a particular sense of a predicate name. A term t is either a constant or a variable. The symbol $\oplus$ denotes sequence concatenation.

(25) The meaning of a constant, relative to an input and output interpretation function and to ${}^{I}[\![c_m]\!]^{I,\langle s,g,w\rangle} = I(c, s(m))$, iff $s(m)$ is defined, otherwise, ${}^{I}[\![c]\!]^{I,\langle s,g,w\rangle} = I(c, w)$.[5]

(26) ${}^{I}[\![x]\!]^{I,\langle s,g,w\rangle} = g(x)$

(27) ${}^{I}[\![\langle t^1, ..., t^n\rangle]\!]^{I,\langle s,g,w\rangle} = \langle {}^{I}[\![t^1]\!]^{I,\langle s,g,w\rangle}, ..., {}^{I}[\![t^n]\!]^{I,\langle s,g,w\rangle}\rangle$

(28) A predication is true if the denotation of its arguments, as a tuple, is in the interpretation of the predicate at the relativized sense. If the tuple is 0-ary, then it is a proposition which is true in the relevant sense if and only if the predicate name and sense pair exist in the general interpretation function (and the proposition is otherwise false.
${}^{I}[\![P^n(\sigma)]\!]^{I,\langle s,g,w\rangle} = 1$ iff $n \geq 0, |\ \sigma\ | = n$ and $\langle P, w\rangle \oplus {}^{I}[\![\sigma]\!]^{I,\langle s,g,w\rangle} \in I$

(29) ${}^{I}[\![\neg P]\!]^{I,\langle s,g,w\rangle} = 1$ iff ${}^{I}[\![P]\!]^{I,\langle s,g,w\rangle} = \emptyset$

(30) ${}^{I}[\![P \wedge Q]\!]^{I,\langle s,g,w\rangle} = 1$ iff ${}^{I}[\![P]\!]^{I,\langle s,g,w\rangle} = 1$ and ${}^{I}[\![Q]\!]^{I,\langle s,g,w\rangle} = 1$

(31) ${}^{I}[\![\forall x \phi]\!]^{I,\langle s,g,w\rangle} = 1$ iff ${}^{I}[\![\phi]\!]^{I,\langle s,g[x/d],w\rangle} = 1$ for each element d of the domain, where $g[x/d]$ is an assignment function just like g apart from the assignment to x, which is instead d.

(32) Existential quantification is interpreted in predictably different metalanguage from (31).

These clauses are static in that the output interpretation is always identical to the input interpretation.

2.3.2. Sense-relative assertion

This section refines the definitions for assertion provided by Vogel (2001). In that proposal, static interpretation was reserved for senses classified as literal and dynamic interpretation for senses classified as non-literal. What is correct about this distinction is that the difference between a literal sense and a non-literal sense is convention in classifying it as such. Here, a partial ordering in that dimension is assumed (this emerges more below, particularly in how this relates to genericity). Evidently, people are able to perceive degrees of metaphoricity (Ortony, 1979). I argue that Vogel (2001) was incorrect in leaving open the suggestion that only non-literal expressions are open to belief revision; the independent need for sense extension and contraction was motivated in §2.1. Assume that δ is an act of deixis or reference designation that may be used to pick out individuals or tuples of individuals. Again, let t be a term, a constant or a variable. In some cases, assertional interpretation is not defined as such, but reduces to static interpretation.

(33) Given an input interpretation function, assertional interpretation of a constant with a particular designation

[4] An article in *The Economist* may use without penalty "bank" in an article reviewing property values on one side of the Seine.

[5] This "otherwise" reference to a default sense is to be assumed consistently throughout the remainder.

of sense may point to an individual referred to with accompanying deixis; the constant refers, and the interpretation function is augmented with an additional tuple, appropriately.

${}^{I}[\![c_m]\!]_{+}^{I\cup\{\langle c, s(m), \delta(c)\rangle\},\langle s,g,w\rangle} = \delta(c)$, iff $\delta(c)$ is defined .

(34) If no designation of sense if supplied, then assertional interpretation of a constant is relative to a default sense,[6] if an individual is also designated, and otherwise assertional interpretation of a constant reduces to static interpretation.

${}^{I}[\![c]\!]_{+}^{I\cup\{\langle c, w, \delta(c)\rangle\},\langle s,g,w\rangle} = \delta(c)$, iff $\delta(c)$ is defined .

(35) ${}^{I}[\![\langle t^1, ..., t^n\rangle]\!]_{+}^{O,\langle s,g,w\rangle} =$

$\langle {}^{I}[\![t^1]\!]_{+}^{O^1,\langle s,g,w\rangle}, ..., {}^{O^{n-1}}[\![t^n]\!]_{+}^{O,\langle s,g,w\rangle}\rangle$

(36) Note that the assertional interpretation of a predication (or proposition) always succeeds relative to either a designated or default sense. It has the effect of adding a tuple (possibly empty for a proposition) to the characteristic function for the n-ary predicate for the relevant sense. The interpretation of constants used as arguments may be extended to new senses denoting new individuals along the way, via (35)

${}^{I}[\![P_m^n(\sigma)]\!]_{+}^{I\cup\left\{\langle P, s(m)\rangle\oplus^{I}[\![\sigma]\!]_{+}^{O,}\langle s,g,w\rangle\right\}\cup O,\langle s,g,w\rangle} = 1$

By construction, the assertional interpretation of (36), if repeated for sufficient designations of elements of the domain, can come to make the static interpretation of the universal quantifier provided in (31) work out to be true, and it can make existential generalizations true in a single application for the relevant sense. While in §2.3.1., static interpretation clauses for implication and disjunction are omitted because they can be defined from negation and implication, omission of clauses here should imply that interpretation is static. That is, there is no direct clause for extending the sense of a predicate under the scope of a quantifier, but doing so with individual constant terms will have the effect of making static interpretation relative to the selected sense work out to be true. On the other hand, senses of predicate names and constants cannot, by this construction, be augmented under the scope of negation. However, because extension of a predicate at an index for a sense provides grounds for static interpretation of an existential generalization to be true, it equally supplies grounds for a formerly true negated existential generalization to be false. Even just addition of truths inside the model yields nomonotonicity in support of sentences in the language.

[6] In general, if a sense is not designated then interpretation reverts to being relative to the default; the same holds for (36), for example.

2.3.3. Sense-relative retraction

Like Lemon (1998), I will assume that names of individuals cannot be retracted. Thus, names and tuples of names will be interpreted as what they mean according to a static designated sense. The output of retracting information about a particular tuple of individuals from the denotation of a predicate for some sense of the predicate is an interpretation function which is smaller (if that tuple was in I for the predicate at that sense in the first place), and the formula will evaluate to be false. Subsequent static interpretation of the negated formula, picking out exactly that same tuple, will evaluate as true because the non-negated form is now false.

(37) ${}^{I}[\![P_m^n(\sigma)]\!]_{\dot{-}}^{I-\left\{\langle P, s(m)\rangle\oplus^{I}[\![\sigma]\!]^{I,}\langle s,g,w\rangle\right\},\langle s,g,w\rangle} = 0$

Universally quantified formulae (possibly complex) may be retracted by deleting a tuple from the interpretation function that creates an exception. Existentially quantified formulae may be retracted by deleting all tuples that support the existential generalization. The only generalization over Lemon's work provided in this section is that retraction of information is relativized to the sense of the predicate at stake. It uses an extensional unpacking of intensions.

2.4. Initial reflections on metaphoricity

The discussion which precedes has not provided the logic which fits the constraints on updating and downdating models as specified. Ensuring the correspondence between alterations to models and closure of the set of sentences true in those models is a separate exercise. However, it can be seen from what is discussed what sentences will gain or lose support and that the entire system is non-monotonic, because the underlying models are non-monotonic: relations can expand and contract. The location of dynamic semantics for the language is in the non-logical expressions—proposition and predicate names as well as names of individuals (all relative to senses of them). It is possible to imagine varying the interpretation of the logical constants ($\wedge, \neg$, etc.) so that they do not behave in classical ways (Kuhn, 1981); however, that is not of focus here. The language is set up such that in NPs, head noun restrictor sets; in VPs, verbal heads; in APs, adjectives and adverbs; in PPs, prepositions may expand and contract the sets that they are true of as individuals or tuples of individuals corresponding to relations.

It is assumed that these sets are the input to generalized quantifier constructions (Barwise and Cooper, 1981) to, for example, construct an NP as a set of sets which "lives on" its head noun set, and such that a sentence involving an NP and an intransitive VP or copula-linked predication is true just if the set given by the predicate is an element of the set of sets provided by the NP. If metaphorical statements are taken to be class inclusion statements, this analysis in terms of generalized quantifiers will demand modification to achieve the same effect. In fact, the inclusion statement is that the "lives on" property holds: whether the characteristic set χ corresponding to any predicate is an element of the quantifier depends only on the intersection of the head

noun set (N) from the quantifier with χ. For any χ that is in the GQ denotation supersets or subsets will either have to also be elements of the GQ donation as well (or must not be) depending on the determiner that combines with the head noun set to form the GQ. Thus, the "lives on" property takes care of class inclusion, but also exclusions where necessary. The reason to accept generalized quantifier theory is its robust account of evidently syntactic puzzles (e.g. the "definiteness effect" in partitive constructions), semantic puzzles (e.g. licensing of negative polarity items by downwards monotone determiners), as well as predicting processing facts about natural language determiners (e.g. monotonic increasing determiners (e.g. "some" and "all") are easier to evaluate than monotone decreasing determiners (e.g. "no" and "few"), which are in turn easier than non-monotonic determiners (e.g. "exactly three")) that are supported by empirical evidence (Moxey and Sanford, 1993). Ample reason to move to a generalize quantifier account are provided by Barwise and Cooper (1981); primary is that first-order logic does not have the expressive capacity to represent the meaning of "counting" as is required by relatively mundane natural language determiners like "most" or "many".[7] Finally, in presenting the invariants associated with generalized quantifiers, Barwise and Cooper (1981) assumed a fixed-model constraint to address the variance in determiner meaning that depends on contextual factors like expectations. For example, a different number of people, even a different proportion of a relevant head noun set being quantified over, might count as "many" depending on the expectations. The fact, that the cardinality or ratio involved in "many" is to be interpreted with varying models in generalized quantifier theory is a background support for the kind of variation in interpretation depending on signalled sense to account for aspects of metaphoricity in this paper. Consider the highlighted portion of (38).[8]

(38) There was never a solicitation for money at these events, but of course, the President hoped that people in this category of friends and prior supporters would give money afterwards. *And, in fact, many did, and many did not.*

It is clear that metaphoricity is handled here by classification of senses of predicates as metaphorical or not, and degrees of metaphoricity can be represented. It remains to discuss more about the nature of the distinct senses of predicates and what makes them stand in special relationships to their base forms. The basic idea is that by addressing predicates and their related senses, one has access to a larger characteristic function for the set than is relevant to any literal sense of the predicate. Each possible sense is the characteristic function corresponding to an abstraction over salient properties associated with the characteristic function for the predicate. There can be any number of such abstractions, and one does not expect each of them to have a unique name (Glucksberg, 2001). Each additional sense of a predicate has its own characteristic function, and as has been seen, the set determined by each such function can be expanded or contracted using the dynamic interpretation mechanisms specified above. Equivalence classes of senses of a predicate form the space of polysemy for a predicate (as distinguished from its having unrelated homonymic senses), and all of the tuples in the entire equivalence class form a larger set than those in the basic literal sense.

3. Metaphoricity and Genericity

As constructed, predicates cannot be extended to cover new tuples under the scope of negation, but negations can be made true by retracting tuples from the characteristic functions of particular senses of predicate names. It is tempting to say that novel use of metaphor involves the generation and population of new senses of predicates; conventionalized metaphor is about the re-use of old senses, and dead metaphor does not even involve extending the predicate to a fresh set of tuples. However, a key point here is that information assertion and retraction about individuals and tuples of individuals is independent of metaphoricity being involved. It happens with literal information also.

If senses associated with predicates are individuated, then it is possible to consider subsets of the interpretation function as bundling predicates together by senses that are shared. For example, there is a financial institution sense of "bank" that is in common with a particular sense of "bond". The two words do not mean the same thing: even relative to that shared sense the words participate in different networks of implications and are true of different tuples. It is possible to partially order names of relations paired with their senses in a cline of metaphoricity. The different senses of predicates will ultimately be true of different sets of tuples. In discussing abstractions that yield senses of predicates and constant names related to metaphoricity, one obtains sets that have more entities in them in their totality than the literal sense that one started with.

Genericity provides an alternative sense to predicates that has nearly identical properties to metaphorical sentences, but on the analysis provide here, they are explained by appeal to construction of related contracted senses of predicates. Like metaphors, generics can be predications over nominals (39)-(41) or can involve the verbs directly as well (42). Generics certainly cannot be understood as universally quantified statements, as their nature is to have exceptions. Thus, if generics are taken to be category inclusion statements, they turn out to be false in their literal sense. However, generics cannot be truthfully understood as asserting even that *most* of the entities in the subject NPs head noun set have the predicated property, because (39)-(41) would remain true if there tend to be more male platypuses than female ones, or even if most platypuses die before reaching the age of being able to reproduce. Similarly, (42)

[7] Note that Glucksberg (2001, pg. 22) recalls experiments from 1982 and 1989 which revealed significant differences in responses to metaphorical statements with quantified subjects depending on the determiner of quantification ("some" vs. "all"); one might anticipate that a wide range of variability is indexed by exactly the monotonicity properties of the determiner.

[8] Attributed to Lanny Davis, special White House counsel, February 25, 1997. OnLine Focus interview with Elizabeth Farnsworth (`http://www.pbs.org/newshour/bb/white_house/february97/davis_2-25.html` — last verified March 5, 2008.

might be uttered to mean that the only time Leslie smokes, it's after dinner, or among the times that Leslie smokes, after dinner times are included. The safest "strong" reading of a generic in first-order languages is that the sentences make an existential claim that, for example, there is at least one platypus that has produced an egg. However, the existential readings are a challenge for sentences like (43) in which there is no real entity in the domain that satisfies the existential generalization, but perhaps appeal to fictional entities could provide some sort of straw to grasp to make it work.

(39) The platypus is an egg laying mammal.

(40) A platypus is an egg laying mammal.

(41) Platypuses are egg laying mammals.

(42) Leslie smokes after dinner.

(43) Unicorns are white.

(44) An egg laying mammal is the platypus.

The truth conditions of generics are thus as troubled as those of metaphors. Note further that reversing the predications is possible, but changes the meaning slightly, admitting a Gricean implicature in (44) that there are other egg laying mammals as well. This reversibility issue is comparable to the situation referred to above (6) and (7) with metaphors.

It is common to understand generics as involving a restricted domain of quantification over salient individuals. This is rather the converse of what happens with metaphor understanding. Thus, the proposal to unify the treatment of metaphoricity and genericity in this dynamic framework is to allow for alternative senses of literal predicates which are reduced by individuals or tuples[9] that challenge the literal truth of universal quantification over the full domain. Metaphors are class inclusion statements that involve expanding hitherto un-named categories, and generics are class inclusion statements that involve shrinking categories with prior names. Among the alternative senses for predicates are those which stand systematically in this way via relevant restriction over the characteristic set of the predicate at some sense.

4. The framework in light of Glucksberg

One aspect of the system that merits discussion is its main area of divergence from the work of Glucksberg and colleagues. This is with respect to the question of asymmetry of metaphor, which I argued above extends somewhat to genericity. The divergence is in that the system doesn't place great emphasis on the asymmetry beyond the order of arguments in a tuple, which is in each case an ordered sequence. The system, through multiplicity of senses for predicates and terms, admits duality of reference, but it is not prejudiced to require that the dual argument must be in a non-subject position. Interestingly, Glucksberg (2001) comments in a number of places less on the asymmetry of subject and object, as with respect to new and given. This is also called the topic-comment distinction, and it often in English coincides with the grammatical subject, but it is not analytically identical (Keenan, 1975).

(45) Einstein [my brother points at a clever companion] can work out how the remote control works.

(46) It is sharks that lawyers are.

(47) Sharks, Lawyers are.

First of all, (45) shows that the Demjanjuk examples of Glucksberg (2001, pg. 40) involving abstract categories can occur in subject position. The cleft (46) and topicalization (47) are both constructions that move canonical objects into a topic position for information packaging purposes, and in these cases it turns out to be the abstract category that form topic, and the finite sentence with an object gap that forms a predication for the comment. Perhaps one would want to argue that the subject remains given in these and related constructions, but it is clear that it is not the linear order of presentation that matters as much as the information packaging into topic and comment.

However, a more robust class of examples of non-literal expressions best understood as class inclusion statements, but with the class in the initial position, has an exemplar in (48).[10] This construction relates directly to predication metaphor (49). A counterpart construction for simile is perhaps anomalous (51).

(48) "Anyone who has lived in the ethnic shouting match that is New York City knows exactly what I mean"

(49) New York City is an ethnic shouting match.

(50) Anyone who has lived in the New York City that is an ethnic shouting match knows exactly what I mean.

(51) the jail that is like Sandy's job

In (48) both terms of the predication can be understood via literal referent or as concepts, but there is evidently a preference for "the ethnic shouting match" to be understood as a name for category which is asserted to have the literal New York City within it. The relevant nonliteral constituent of (48) can be equally understood via (49). An adapted formulation is provided in (50) to show that reversibility does obtain and "New York City" does not appear to be forced into a sub-kind level expression, although it has to be at least a category here for the definite reference to work. The point is that there is more to explore about the asymmetry facts associated with metaphor. They appear to be not simply about the order of presentation of topic and vehicle and their reversibility. The facts seem to depend upon the construction which is used to package the relevant information. In the system provided in this paper, (35) gives the dynamic interpretation of terms in a tuple, interpretation of the output of the first as the input to the second, and so on. The tuples are ordered by the argument structure of the predicate, rather than the information packaging of the construction it appears in. There may well be empirical

[9] Individuals are singleton tuples, anyway.

[10] Attributed to Andrew Sullivan by Roberts (2007).

consequences that depend on alternative information packaging associated with argument terms, but it is not clear that they have much significance. That is, while a tendency to restrict reversibility of arguments and correlation with topic-comment structures may be useful diagnostics of metaphoricity, the dual reference theory seems to be able to stand up independently in cases where the data seems slightly at odds with the asymmetry claims.

5. Final Remarks

This paper has argued that metaphoricity and genericity are best handled within the same semantic framework, one that admits information update, names of individuals and predications paired with senses. The formal machinery has been sketched in an extensional unpacking of the main ideas. Pairs of predicate names and senses can be partially ordered to achieve a continuum of metaphoricity. Glucksberg (2001) has argued that metaphors are best analyzed as class inclusion statements involving dual reference. Generics and habituals certainly look like class inclusion statements and show many of the same properties of non-literal interpretation that metaphors do. It has been shown exactly how metaphors relate to each other within a non-monotonic system for information change.

6. Acknowledgements

This research is supported by Science Foundation Ireland RFP 05/RF/CMS002. I am grateful for constructive feedback from the reviewers, but I fear that my adjustments so far in light of their suggestions do not do them justice.

7. References

C. E. Alchourrón, P. Gärdenfors, and D. Makinson. 1985. On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Journal of Symbolic Logic*, 50:510–30.

Jon Barwise and Robin Cooper. 1981. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, 4(2):159–219.

Greg Carlson and Jeff Pelletier, editors. 1995. *The Generic Book*. University of Chicago Press.

Ariel Cohen. 1999. *Think Generic! The Meaning and Use of Generic Sentences*. Stanford: CSLI Publications.

Ariel Cohen. 2001. On the Generic Use of Indefinite Singulars. *Journal of Semantics*, 18(3):183–209.

Donald Davidson. 1984. What Metaphors Mean. In Donald Davidson, editor, *Inquiries into Truth and Interpretation*, pages 245–64. Oxford: Oxford University Press.

L.T.F. Gamut. 1991. *Language, Logic and Meaning, Part 1: Introduction to Logic*. Chicago University Press, Chicago.

Sheila Glasbey. 2007. Aspectual Composition in Idioms. In Louise de Saussure, Jacques Moeschler, and Genoveva Puskas, editors, *Recent Advances in the Syntax and Semantics of Tense, Aspect and Modality*, pages 1–15. Berlin: Mouton de Gruyter.

Sam Glucksberg and Boaz Keysar. 1993. How Metaphors Work. In Andrew Ortony, editor, *Metaphor and Thought*, pages 357–400. Cambridge University Press, 2nd edition. First Published, 1979.

Sam Glucksberg. 2001. *Understanding Figurative Language: From Metphors to Idioms*. Oxford University Press. With a contribution from Matthew S. McGlone.

Andrew Goatly. 1997. *The Language of Metaphors*. Routledge.

J. Groenendijk and M. Stokhof. 1991. Dynamic Predicate Logic. *Linguistics and Philosophy*, 14:39–100.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer.

Edward Keenan. 1975. Towards a Universal Definition of 'Subject'. In Charles Li, editor, *Subject and Topic*, pages 304–33. Academic Press: London. Symposium on Subject and Topic, University of California, Santa Barbara, 1975.

Manfred Krifka, Francis Jeffry Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link, and Gennaro Chierchia. 1995. Genericity: An Introduction. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*. University of Chicago Press.

Stephen Kuhn. 1981. Operator Logic. *The Journal of Philosophy*, 78(9):487–499.

Oliver Lemon. 1998. First-order Theory Change Systems and their Dynamic Semantics. In J. Ginzburg, Z. Khasidashvili, C. Vogel, J-J. Levy, and E. Vallduvi, editors, *The Tbilisi Symposium on Logic, Language and Computation: Selected Papers*, pages 85–99. SiLLI/CSLI Publications.

Linda M. Moxey and Anthony J. Sanford. 1993. *Communicating Quantities: A Psychological Perspective*. Lawrence Erlbaum.

Andrew Ortony. 1979. Beyond Literal Similarity. *Psychological Review*, 86(3):161–180.

Sam Roberts. 2007. Podcast: The New York Brand of Bias. *The New York Times*. `http://cityroom.blogs.nytimes.com/2007/07/26/podcast-the-new-york-brand-of-bias/`— last verified, April 2008.

Josef Van Genabith. 2001. Metaphor, Logic and Type Theory. *Metaphor and Symbol*, 16(1 & 2):43–57.

Henk J. Verkuyl. 1993. *A Theory of Aspectuality*. Cambridge Studies in Linguistics. Cambridge University Press.

Carl Vogel and Michelle McGillion. 2002. Genericity is Conceptual, not Semantic. In Gábor Alberti, Kata Balogh, and Paul Dekker, editors, *Proceedings of the Seventh Symposium on Logic and Language*, pages 163–72. 26-29 August, 2002, Pécs, Hungary.

Carl Vogel. 2001. Dynamic Semantics for Metaphor. *Metaphor and Symbol*, 16(1 & 2):59–74.

Method

POSTER

Revisiting the Use of Lexically–Based Features for Sentiment Detection

Ben Allison
University of Sheffield
ben@dcs.shef.ac.uk

Abstract

This paper addresses the problem of supervised sentiment detection using classifiers which are derived from word features. We argue that, while the literature has suggested the use of lexical features is inappropriate for sentiment detection, a careful and thorough evaluation reveals a less clear–cut state of affairs. We present results from five classifiers using word based–features on three tasks, and show that the variation between classifiers can often be as great as has been reported between different feature sets with a fixed classifier. We are thus led to conclude that classifier choice plays at least as important a role as feature choice, and that in many cases word–based classifiers perform well on the sentiment detection task.

1. Introduction

Sentiment detection as we approach it in this paper is the task of ascribing one of a pre– and (well–) defined set of non–overlapping sentiment labels to a document. Approached in this way, the problem has received some considerable attention in recent computational linguistics literature, and early references are (Pang et al., 2002; Turney and Littman, 2003).

Whilst it is by no means obligatory, posed in such a way the problem can easily be approached as one of classification. The precise nature of the classification problem depends upon its particulars – if training data for the sentiments of interest are available, it can be approached as a supervised problem (which is relatively well defined); if they are not, it is unsupervised. The unsupervised scenario poses something of a problem since sentiment is unlikely to be the sole characteristic of a text and in many ways will be secondary to more obvious dimensions (such as topic and author) which one would expect to be more readily captured by any unsupervised partitioning of the texts. For this reason, for the purposes of this paper we restrict our attentions to the former problem.

Within the scope of the supervised classification problem, to use standard machine learning techniques one must make a decision about the features one wishes to use – that is, one must decide how the texts are to be reduced to numeric values. Typically a single text is represented as a vector, and it is common to refer to each of the elements of the vector individually, and the process or aspect of the text to which it relates, as a feature.

Several authors have remarked that for sentiment classification, lexically–based features (that is, features which describe the frequency of use of some word or combination of words, or some transform thereof) are generally unsuitable for the purposes of sentiment classification. For example, (Efron, 2004) bemoans the "initially dismal word–based performance", and (Mullen and Malouf, 2006) conclude their work by saying that "traditional word–based text classification methods (are) inadequate" for the variant of sentiment detection they approach.

This paper revisits the problem of supervised sentiment detection, and whether lexically–based features are adequate for the task in hand. We conclude that, far from providing overwhelming evidence supporting the previous position, an extensive and careful evaluation leads to generally good performance on a range of tasks. However, it emerges that the choice of method plays at least as large a role in the eventual performance as is often claimed for differing representations and feature sets. We suggest that there can be no general conclusion about the performance of word–based features, and the performance of a feature set depends heavily upon the method with which it is used. We cannot therefore help but conclude that in certain conditions, lexically–based classifiers show themselves to be well–suited to the sentiment detection task, and furthermore can be deployed without the need for empirical calibration to the particular task in hand.

The rest of this paper is organised as follows: §2. describes our evaluation in detail; §3. describes the classifiers we use for these experiments; §4. presents results and informally describes trends, which are supported by a more formal analysis in §5.. Finally, §6. ends with some brief concluding remarks.

2. Experimental Setup

The evaluation presented in this work is on the basis of three tasks: the first two are the movie review collection first presented in (Pang et al., 2002) which has received a great deal of attention in the literature since, and the collection of political speeches presented in (Thomas et al., 2006). Since both of these data sets are binary (i.e. two-way) classification problems, we also consider third problem, using a new corpus which continues the political theme but includes five classes. Each of them is described separately below.

The movie review task is to determine the sentiment of the author of a review towards the film he is reviewing – a review is either positive or negative. We use version 2.0 of the movie review data.[1]

The task for the political speech data is to determine whether an utterance is in support of a motion, or in opposition to it, and the source of the data is automatically

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/

transcribed political debates. For this work, we use version 1.1 of the political data.[2]

The new collection consists of text taken from the election manifestos of five UK political parties for the last three general elections (that is, for the elections in 1997, 2001 and 2005). The parties used were: Labour, Conservative, Liberal Democrat, the British National Party and Sinn Féin. As such, the data represent a broad spectrum of political opinion, from moderate conservative and liberal parties, as well as the extreme right–wing British National Party, and Sinn Féin, which is reported to have been linked to the IRA. The corpus is approximately 250,000 words in total, and we divide the manifestos into "documents" by selecting non–overlapping twenty–sentence sections. This results in a corpus of approximately 650 documents, each of which is roughly 300—400 words in length.

The following is an example of the BNP's manifesto, and illustrates their extreme policies (albeit cloaked in relatively innocuous language):

> To ensure that the British people retain their homeland and identity, we call for an immediate halt to all further non-white immigration, the immediate deportation of criminal and illegal immigrants, and the introduction of a system of voluntary resettlement whereby those non-white immigrants who are legally here will be encouraged, but not compelled, to return to their lands of ethnic origin. The BNP will ensure the proper financial and material provisions are available for this, with the aim of halting and reversing the trend towards a non-white Britain, and ensuring that the British people have a homeland and retain their unique identity.

In contrast, Sinn Féin's politics are clear from sections of their manifesto:

> The primary political objectives of Sinn Féin are Irish unity, political independence, sovereignty and national reconciliation. We are working to achieve this in our lifetime.
>
> Sinn Féin has consistently urged an island-wide approach in key policy areas. We have given practical expression to this through the work of our ministers in the Executive, the all-Ireland Ministerial Council, in Leinster House, the Assembly and the European Parliament.

We also wished to test the impact of the amount of training data; various studies have shown this to be an important consideration when evaluating classification methods. Of particular relevance to our work and results is that of (Banko and Brill, 2001), who show that the relative performances of different methods changes as the amount of training data increases. Thus we vary the percentage of documents used as training between 10% and 90% at 10% increments. For a fixed percentage level, we select that percentage of documents from each class (thus maintaining class distribution) randomly as training, and use all remaining as testing. We repeat this procedure five times for each percentage level. All results are in terms of the simplest performance measure, and that most frequently used for non–overlapping classification problems, accuracy.

Otherwise, all "words" are identified as contiguous alpha–numeric strings. We use no stemming, no stoplisting, no feature selection and no minimum–frequency cutoff.

We were also interested to observe the effects of restricting the vocabulary of texts to contain only words with some emotional significance, since this in some ways seems a natural strategy, ignoring words with specific topical and authorial associations. We thus perform experiments on the movie review collection, but using only words which are marked as *Positive* or *Negative* in the General Inquirer Dictionary (Stone et al., 1966).

3. Methods

This section describes the methods we evaluate in detail. To test the applicability of both word–presence features and word–count features, we include standard probabilistic methods designed specifically for these representations. We also include a more advanced probabilistic method with two possibilities for parameter estimation, and finally we test an SVM classifier, which is something of a standard in the literature.

3.1. Probabilistic Methods

In this section, we briefly describe the use of a model of language as applied to the problem of document classification, and also how we estimate all relevant parameters for the work which follows.

We consider cases where documents are represented as vectors of count–valued (possibly only zero or one, in the case of binary features) random variables such that $d = \{d_1 ... d_v\}$. As with most other work, we will further assume that the words are in a document are exchangeable and hence can be represented simply by the number of times each word occurs.

In classification, interest centres on the posterior distribution of the class variable, given a document. Where documents are to be assigned to one class only (as in the case of this paper), this class is judged to be the most probable class.

Classifiers such as those considered here model the posterior distribution of interest from the joint distribution of class and document. This means incorporating a *sampling model*, which encodes assumptions about how documents are sampled. Thus letting $\tilde{c}$ be a random variable representing class and $\tilde{d}$ be a random variable representing a document, the estimate is:

$$\Pr(c|d) \propto \Pr(c) \cdot \Pr(d|c) \tag{1}$$

Which can be normalised by including the factor $\frac{1}{\Pr(d)}$. However, since this factor does not depend on the class, it can be ignored if the goal is to find the most probable c. For the purposes of this work we also assume a uniform prior on c, meaning the ultimate decision is on the basis of the document alone.

[2] http://www.cs.cornell.edu/home/llee/data/convote.html

For each of the probabilistic methods, we describe the relevant distributions and how parameters are estimated for a *fixed* class. We estimate a single model of the types shown below for each possible class, and combine estimates to make a decision as above, and as such we will drop subscripts referring to a particular class for clarity in notation. Similarly, where training documents and/or counts are mentioned, these relate only to the class in question.

Binary Independence Sampling Model

For a vocabulary with v distinct types, the simplest representation of a document is as a vector of length v, where each element of the vector corresponds to a particular word and may take on either of two values: 1, indicating that the word appears in the document, and 0, indicating that it does not. Such a scheme a long heritage in information retrieval: see e.g. (Lewis, 1998) for a survey, and (Robertson and Jones, 1988; McCallum and Nigam, 1998) for applications in information retrieval and classification respectively. This model depends upon parameter θ, which is a vector also of length v, representing the probabilities that each of the v words is used in a document.

Given these parameters (and further assuming independence between components of d), the term $p(d|c)$ is simply the product of the probabilities of each of the random variables taking on the value that they do. Thus the probability that the j-th component of $\tilde{d}$, $\tilde{d}_j$ is one is simply θ_j (the probability that it is zero is just $1 - \theta_j$) and the probability of the whole vector is:

$$p_{bin-indep}(d|\theta) = \prod_j p_{bi}(d_j|\theta_i) \tag{2}$$

where:

$$p_{bi}(d_j|\theta_j) \quad = \quad \begin{cases} \theta_j & \text{if } d_j = 1 \\ 1 - \theta_j & \text{otherwise} \end{cases} \tag{3}$$

In a slight departure from previous work, we assume each of the θ_j is the parameter to a binomial distribution. Then (taking the Bayesian approach to estimation) given a set of k training documents $\mathcal{D}$ with values of d_j $\mathcal{D}_j = (d_{1j}...d_{kj})$ and assuming a prior uniform on $[0, 1]$, θ_j has posterior distribution which is $Beta(1 + \sum_i d_{ij},\ 1 + k - \sum d_{ij})$. The expected value of the posterior (and thus the estimate for θ_j) is then:

$$\hat{\theta_j} = \frac{1 + \sum_i d_{ij}}{2 + k} \tag{4}$$

Which is equivalent to a maximum likelihood estimate if one supplements the actual training documents with two pseudo-documents, one in which every word occurs and one in which none of the words occur (simple maximum likelihood estimates alone do not suffice, since there is the possibility that $\hat{\theta_j} = 0$ or 1, which would collapse the whole calculation if word j were to occur or not occur, respectively).

Multinomial Sampling Model

A natural way to model the distribution of word *counts* (rather than the presence or absence of words) is to let $p(d|c)$ be distributed multinomially, as proposed in (Guthrie et al., 1994; McCallum and Nigam, 1998) amongst others. The multinomial model assumes that documents are the result of repeated trials, where on each trial a word is selected at random, and the probability of selecting the j-th word is θ_j.

Using multinomial sampling, the term $p(d|c)$ has distribution:

$$p_{multinomial}(d|\theta) = \frac{\left(\sum_j d_j\right)!}{\prod_j (d_j!)} \prod_j \theta_j^{d_j} \tag{5}$$

A simple Bayes estimator for θ can be obtained by taking the prior for θ as a Dirichlet distribution, in which case the unnormalised posterior is also Dirichlet. Denote the total training data for the class in question as $\mathcal{D} = \{(d_{11}...d_{1v})\,...\,(d_{k1}...d_{kv})\}$ (again, there are k training documents each of which has words counts for each of v words). If $p(\theta) \sim Dirichlet(\alpha_1...\alpha_v)$, then the mean of $p(\theta|\mathcal{D})$ for the j-th component of θ (which is the estimate we use) is:

$$\hat{\theta_j} = \mathrm{E}[\theta_j|\mathcal{D}] = \frac{\alpha_j + n_j}{\sum_j \alpha_j + n_\bullet} \tag{6}$$

where the n_j are the sufficient statistics $\sum_i d_{ij}$, and $n_\bullet$ is $\sum_j n_j$. We follow common practice and use a standard reference Dirichlet prior, such that $\alpha_j = 1$ for all j.

3.1.1. Hierarchical Sampling Models

In contrast to the model above, a hierarchical sampling model assumes that $\tilde{\theta}$ varies between documents, and has distribution which depends upon parameters η. This allows for a more realistic model, assuming that the probabilities of using words vary between documents, and are only subject to some general trend.

For example, consider documents about politics: some will discuss the current British Prime Minister, Gordon Brown. In these documents, the probability of using the word *brown* (assuming case normalisation) may be relatively high – perhaps as much as $\frac{1}{100}$. However, other politics articles may discuss US politics, for example, or the UN, French elections, and so on, and these articles may have a much lower probability of using the word *brown*, say $\frac{1}{10000}$: in these cases there may be just the occasional reference to the Prime Minister. This discussion is something of a simplification, since the true model hypothesises that the count of the word *brown* in each document depends upon a different θ_j; nevertheless, the example captures some of the intuition of the model.

Starting with the joint distribution $p(\theta, d|\eta)$ and averaging over all possible values that θ may take in the new document gives:

$$p(d|\eta) = \int p(\theta|\eta)p(d|\theta)\,d\theta \tag{7}$$

where integration is understood to be over the entire range of possible θ. Intuitively, this allows $\tilde{\theta}$ to vary between documents subject to the restriction that $\tilde{\theta} \sim p(\theta|\eta)$, and the probability of observing a document is the average of its

probability for all possible θ, weighted by $p(\theta|\eta)$. The generative model is such that θ is first sampled from $p(\theta|\eta)$ and then d is sampled from $p(d|\theta)$, leading to the hierarchical name for such models.

A Joint Beta-Binomial Sampling Model

There are several possible instantiations of the general scheme above, and (Madsen et al., 2005) provide an example. However, there are certain theoretical issues to be addressed with that work, and so we use an alternate model within the framework outlined above.

We propose to decompose the term $p(d|\eta)$ into a sequence of independent terms of the form $p(d_j|\eta_j)$. A natural way for each of these terms to be distributed is to let the probability $p(d_j|\theta_j)$ be binomial (as it is if $p(d|\theta)$ is multinomial) and to let $p(\theta_j|\eta_j)$ be beta–distributed. The probability $p(d_j|\eta_j)$ (where $\eta_j = \{\alpha_j, \beta_j\}$, the parameters of the beta distribution) is then:

$$p_{bb}(d_j|\alpha_j,\beta_j) = \frac{n!}{d_j!(n-d_j)!} \times \frac{\mathrm{B}(d_j+\alpha_j,\, n-d_j+\beta_j)}{\mathrm{B}(\alpha_j,\beta_j)} \quad (8)$$

where $\mathrm{B}(\bullet)$ is the Beta function. The term $p(d|\eta)$ is then simply:

$$p_{beta-binomial}(d|\eta) = \prod_j p(d_j|\eta_j) \quad (9)$$

This allows means and variances for each of the θ_j to be specified separately, but this comes at a price: the model above does not ensure the sum of parameters is unity. Thus the model is only an approximation to a true model where components of θ have independent means and variances, and the requirements of the multinomial are fulfilled. However, given the inflexibility of other proposed models, we believe such a sacrifice is justified.

As with most previous work, our first estimate of parameters of the beta-binomial model are in closed form, using the method–of–moments estimate proposed in (Jansche, 2003). Method of moments estimates match moments of the sample with moments of the distribution, and solve for unknown parameters. If the count of a word in a document has the distribution above, its expected value is:

$$\mathrm{E}[d_j] = n \times \frac{\alpha_j}{\alpha_j+\beta_j} \quad (10)$$

and its variance is:

$$\mathrm{Var}[d_j] = \frac{n\,\alpha_j\beta_j\,(n+\alpha_j+\beta_j)}{(\alpha_j+\beta_j)^2\,(1+\alpha_j+\beta_j)} \quad (11)$$

Again denoting training data for the j-th word $\mathcal{D}_j = \{d_{1j}...d_{kj}\}$ the theoretical expected value of a sample is simply $\sum_i \mathrm{E}[d_{ij}]$, while the observed expected value is $\sum_i d_{ij}$. Similarly, the theoretical variance of the sample is $\sum_i \mathrm{Var}[d_{ij}]$ while its observed variance is $\sum_i (d_{ij} - \mathrm{E}[d_{ij}])^2$. These are not maximum likelihood estimates, as used in (Lowe, 1999)), but provide a practical estimate of parameters which is feasible for problems with vocabularies with tens of thousands of words, each of which must be modelled for every class, whereas using numeric techniques to find maximum likelihood estimates arguably is not.

This distribution is undefined if any of the α_j are zero. For simplicity, to avoid this we supplement actual training documents with a pseudo-document in which every word occurs once. While not performing the role of a true Bayesian prior, this shares many properties with such an estimate and none of the computational burden.

We also experiment with an alternate estimate, corrected so that documents have the same impact upon parameter estimates regardless of their length. We refer to the original as the Beta–Binomial model, and the modified version as the Alternate Beta-Binomial.

3.2. A Support Vector Machine Classifier

We also experiment with a linear Support Vector Machine, shown in several comparative studies to be the best performing classifier for document categorization (Dumais et al., 1998; Yang and Liu, 1999). Briefly, the support vector machine seeks the hyperplane which maximises the separation between two classes while minimising the magnitude of errors committed by this hyperplane. The preceding goal is posed as an optimization problem, evaluated purely in terms of dot products between the vectors representing individual instances. The flexibility of the machine arises from the possibility to use a whole range of kernel functions, $\phi(x_1, x_2)$ as the result of the dot product in some transformed space.

Despite the apparent flexibility, the majority of NLP work uses the linear kernel such that $\phi(x_1, x_2) = x_1 \cdot x_2$, i.e. there is no transformation. Nevertheless, the linear SVM has been shown to perform extremely well, and so we present results using the the linear kernel from the SVM^{light} toolkit (Joachims, 1999). We use the most typical method for transforming the SVM into a multi-class classifier, the One-Vs-All method, shown to perform extremely competitively (Rennie and Rifkin, 2001).

4. Results

This section presents the results of our experiments on the collections described in §2. The charts present the accuracy for each of the methods as the amount of training data varies, across each of the collections. This section highlights interesting trends in the results; we defer detailed probabilistic analysis of the results to the next section. Also, please note that scales on the y-axes of the following figures changes between charts so as to better illustrate performance differences on the scale appropriate for each collection individually.

Figure 1 shows performance on (Pang et al., 2002)'s movie reviews collection. Several trends are obvious; the first is that, reassuringly, performance generally increases as the amount of training data increases. Note, however, that this is not always the case – a product of the random nature of the training/testing selection process, despite performing the procedure multiple times for each data point. Note also that individual classifiers experience difficulties with particular splits of the data which are not experienced by all.

The most telling example of this is the pronounced dip in the performance of the SVM at 40% training not reflected in other classifiers' performance, despite the fact that all saw the same set of documents both for training and testing. Also, we note that the classifier specifically designed to model binary representations fails to perform as well as the multinomial and Beta–Binomial models – this is in contradiction to (Pang et al., 2002), who observed superior performance using binary features, but inkeeping with results on more standard text classification tasks (McCallum and Nigam, 1998; Jansche, 2003). This once again highlights the danger of basing conclusions on incomplete evaluation.

Figure 2 shows results on the same data using only words marked as positive or negative in the General Inquirer Dictionary. Note here that relative performance trends are markedly different, with the SVM experiencing a particular reversal of fortunes compared to the first figure. Otherwise, the same idiosyncrasies are evident – occasional dips in one classifier's performance not observed with others, and crossing of lines in the graphs.

Figure 3 presents a slightly less changeable picture, although what is apparent is the complete reversal in fortunes of the methods when compared to the previous collection. The binary classifier performs worst by some margin, and the alternate Beta-Binomial classifier is superior by a similar margin. Also, note that at certain points performance for some classifiers dips, while for others it merely plateaus – again, it is important to stress that all classifiers are seeing exactly the same training and testing data.

Finally, Figure 4 displays results from (Thomas et al., 2006)'s collection of political debates. The results here are perhaps the most volatile of all – the impact of using any particular certain classifiers over others is quite pronounced, and the SVM is inferior to the best method by up to 7% in some places. Furthermore, the binary classifier is even worse, and this is exactly the combination used in the original study. The difference between classifiers is in many cases the same as the difference between the general document–based classifier and the modified scheme presented in that paper.

5. Analysis

This section presents a slightly more formal analysis of the significance of some of the points noted in the previous section. The purpose is to establish that several key points are indeed genuine trends rather than random fluctuations, at least on the basis of the chosen data. Where we wish to argue in terms of the significance of the results, we use the following method: we argue that what is of interest is the probability that the true accuracy of one classifier is greater than the true accuracy of another. The true accuracy is the proportion of correctly classified documents we would observe from an unlimited number of test documents, which for two classifiers A and B we will call π_A and π_B.

This can be calculated as follows: suppose we observe classifiers A and B which achieve x_A and x_B correctly classified documents from n possible respectively (although there is strictly no need to assume that the n is the same in both cases, for the purposes of this work this will always be the case, since the classifiers are evaluated over the same sets of documents).

If we assume that, before we observe any results, we are agnostic as to the accuracy we are likely to observe, then our prior distributions on π_A and π_B are uniform, that is all accuracies are as probable as all others *a priori*. This can be made explicit by prior distributions on accuracies which are $Beta(1,1)$ (that is, uniform). If this is the case, then the posterior distributions of accuracies are $Beta(1+x_A, 1+n-x_A)$ and $Beta(1+x_B, 1+n-x_B)$ respectively. The posterior distribution encodes uncertainty in the true accuracy after a limited sample has been observed – thus if $\pi_1 \sim Beta(1,1)$ and $\pi_2 \sim Beta(50,50)$, both have the same expected value of 0.5. However, their variances are dramatically different, and thus we are much more sure that π_2 is close to 0.5 than we are with π_1.

Given these distributions, we calculate $\Pr(\pi_A > \pi_B)$ as the proportion of pairs (π_A, π_B) where π_A and π_B are randomly drawn from their respective posterior distributions, and where $\pi_A > \pi_B$. All probabilities reported here are estimated from 1,000,000 simulated pairs, and are stable to within several decimal places upon repeated runs.

We note first how the amount of training data has an impact on the relative performances in Figure 1. Where we use 60% training (similar to (Pang et al., 2002)'s 3-fold cross validation), we note that the probability that the true accuracy of the multi–variate Bernoulli classifier is greater than that of the SVM is approximately 0.938; that is, we are almost certain this is true. However, for 90% training (the same as would be used for 10–fold cross validation), the two classifiers are inseparable – $\Pr(\pi_{mvb} > \pi_{svm}) \approx 0.5$.

We note also that "improvements" in performances gained by using the General Inquirer for one classifier are anything but for another. For the multinomial classifier, the probability that the unconstrained vocabulary leads to a classifier whose true accuracy is better than that of the same classifier using a constrained vocabulary is approximately 0.976; however, for the SVM this same figure is a less impressive 0.371.

Other trends should need little validation by the procedure above, but we provide some results here for completeness. To demonstrate that increasing the amount of training data improves performance to a significant degree, we note for example that the probability that the multinomial classifier's accuracy is greater with 60% training than with 30% is approximately 0.993 (for the first Movie Reviews collection). The probability that the multinomial (i.e. count–based rather than presence–based features) performs better than the binary independence model at 70% training ≈ 0.962 on the same collection.

We use the same procedure to test other trends noted in the previous section, with similarly conclusive results; we believe this underlines the importance of ruling out simple random fluctuation accounting for apparent variations in performance.

6. Conclusion

In terms of a conclusion, we revisit the initial question. Is it fair to say that the use of lexically–based features leads to classifiers which do not perform acceptably? Of course,

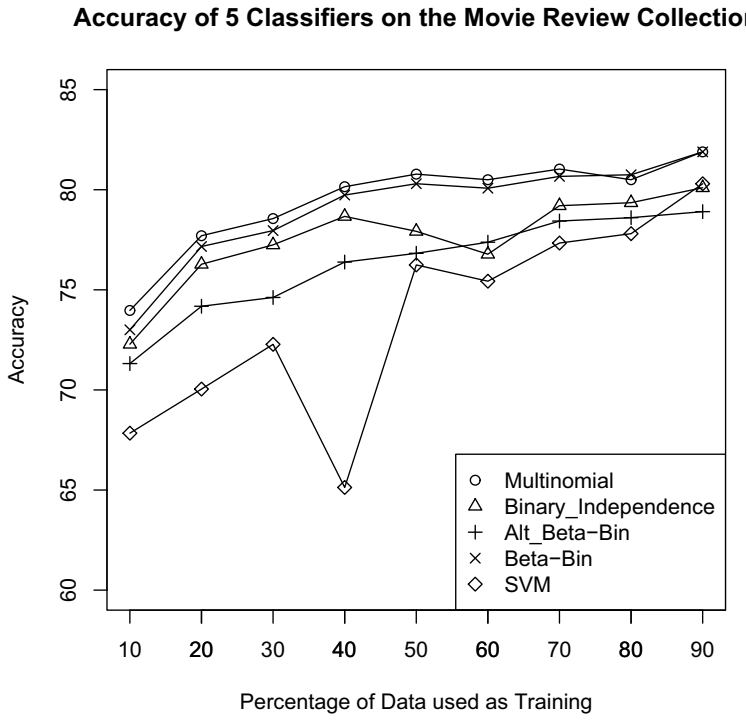

Figure 1: Results for (Pang et al., 2002)'s Movie Review Collection

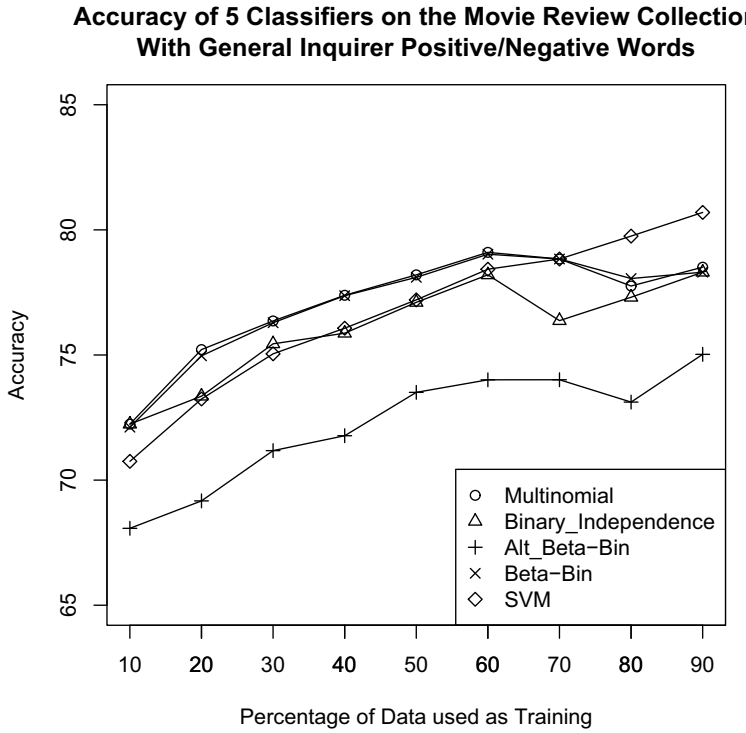

Figure 2: Results for (Pang et al., 2002)'s Movie Review Collection, using only words marked as *Positive* or *Negative* in the General Inquirer Dictionary

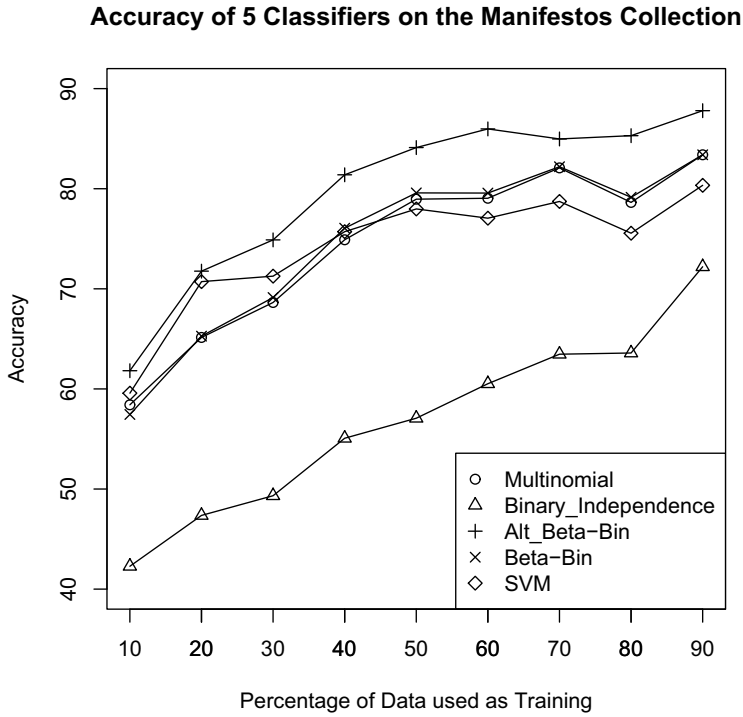

Figure 3: Results for the Manifestos Collection

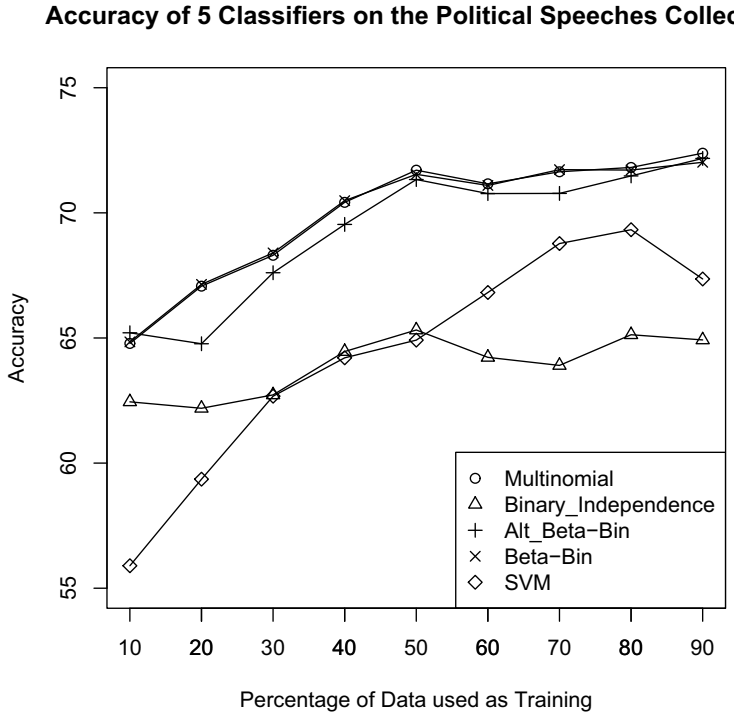

Figure 4: Results for (Thomas et al., 2006)'s Political Speeches Collection

this question glosses over the difficulty of defining "acceptable" performance; however, the only sound answer can be that it depends upon the classifier in question, the amount of training data, and so on. While it would be easier if sweeping generalisations could be made, clearly they are not justified.

Indeed, with the lack of any conclusive evidence, we suggest that one perhaps ought to be strongly drawn to the idea of using lexically–based features for new tasks on the grounds of simplicity. It is well known that complex and over–parametrised models lead to poor generalisation (this is formally encoded in principles such as Occam's Razor, and the practice of Bayesian model comparison (MacKay, 1992)). Certainly, it is hard to argue against the idea that the word–based classifier is the simplest, and in light of results presented here is in certain incarnations comparable with work which uses features much more heavily customised for particular sub–tasks.

7. References

M. Banko and E. Brill. 2001. Mitigating the paucity of data problem: Exploring the effect of training corpus size on classifier performance for nlp. In *Proceedings of the Conference on Human Language Technology*.

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *CIKM '98*, pages 148–155.

Miles Efron. 2004. Cultural orientation: Classifying subjective documents by cociation (sic) analysis. In *Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, pages 41–48.

Louise Guthrie, Elbert Walker, and Joe Guthrie. 1994. Document classification by machine: theory and practice. In *Proceedings COLING '94*, pages 1059–1063.

Martin Jansche. 2003. Parametric models of linguistic count data. In *ACL '03*, pages 288–295.

Thorsten Joachims. 1999. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning*.

David D. Lewis. 1998. Naïve (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98*, pages 4–15.

S. Lowe. 1999. The beta-binomial mixture model and its application to tdt tracking and detection. In *Proceedings of the DARPA Broadcast News Workshop*.

D. J. C. MacKay. 1992. Bayesian model comparison and backprop nets. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 839–846.

Rasmus E. Madsen, David Kauchak, and Charles Elkan. 2005. Modeling word burstiness using the Dirichlet distribution. In *ICML '05*, pages 545–552.

A. McCallum and K. Nigam. 1998. A comparison of event models for naïve bayes text classification. In *Proceedings AAAI-98 Workshop on Learning for Text Categorization*.

Tony Mullen and Robert Malouf. 2006. A preliminary investigation into sentiment analysis for informal political discourse. In *Proceedings of the AAAI Workshop on Analysis of Weblogs*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jason D. M. Rennie and Ryan Rifkin. 2001. Improving multiclass text classification with the Support Vector Machine. Technical report, Massachusetts Insititute of Technology, Artificial Intelligence Laboratory.

Stephen E. Robertson and Karen Sparck Jones. 1988. Relevance weighting of search terms. *Document retrieval systems*, pages 143–160.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.

P. Turney and M. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In *22nd Annual International SIGIR*, pages 42–49, Berkley, August.

Sentiment analysis using automatically labelled financial news

Michel Généreux[1], Thierry Poibeau[2], Moshe Koppel

Laboratoire d'informatique de Paris-Nord – Université Paris 13[1,2], Department of Computer Science – Bar-Ilan University
99 avenue Jean-Baptiste Clément – 93430 Villetaneuse – France[1,2], 52900 Ramat-Gan Israel
{genereux,poibeau}@lipn.univ-paris13.fr, koppel@cs.biu.ac.il

Abstract

Given a corpus of financial news labelled according to the market reaction following their publication, we investigate cotemporeneous and forward-looking price stock movements. Our approach is to provide a pool of relevant textual features to a machine learning algorithm to detect substantial stock price variations. Our two working hypotheses are that the market reaction to a news is a good indicator for labelling financial news, and that a machine learning algorithm can be trained on those news to build models detecting price movement effectively.

1. Introduction

The aim of this research is to build on work by (Koppel and Shtrimberg, 2004) and (V. Lavrenko and Allan, 2000) to investigate the subjective use of language in financial news about companies traded publicly and validate an automated labelling method. More precisely, we are interested in the short-term impact of financial news on the stock price of companies. This is a challenging task because although investors, to a certain extent, make their decision on the basis of factual information such as income statement, cash-flow statements or balance sheet analysis, there is an important part of their decision which is based on a subjective evaluation of events surrounding the activities of a company. Traditional Natural Language Processing (NLP) has so far been concerned with the objective use of language. However, the subjective aspect of human language, i.e. sentiment that cannot be directly inferred from a document's propositional content, has recently emerged as the new useful and insightful area of research in NLP (Devitt and Ahmad, 2007; Mishne, 2007). According to (Wilson and Wiebe, 2003), affective states include opinions, beliefs, thoughts, feelings, goal, sentiments, speculations, praise, criticism and judgements, to which we may add attitude (emotion, warning, stance, uncertainty, condition, cognition, intention and evaluation); they are at the core of subjectivity in human language. We treat short financial news about companies as if they were carrying implicit sentiment about future market direction made explicit by the vocabulary employed and investigate how this *sentimental* vocabulary can be automatically extracted from texts and used for classification. There are several reasons why we would want to do this, the most important being the potential of financial gain based on the exploitation of covert sentiment in the news for short-term investment. On a less pragmatic level, going beyond literal meaning in NLP would be of great theoretical interest for language practitioners in general, but most importantly perhaps, it would be of even greater interest for anyone who wishes to get a sense of what are people feelings towards a particular news, topic or concept. To achieve this we must overcome problems of ambiguity and context-dependency. Sentiment classification is often ambiguous (compare *I had an accident*, negative with *I met him by accident*, not negative) and context dependent (*There was a decline*, negative for *finance* but positive for *crimes*).

2. Experiments

Based on previous work in sentiment analysis for domains such as movie reviews and blog posts, this first series of experiments aim at selecting an appropriate set of three key parameters in text classification: feature *type*, *threshold* and *count*. Our goal is to see whether the most suitable combinations usually employed for other domains can be successfully transferred to the financial domain. Our corpus is a subset of the one used in (Koppel and Shtrimberg, 2004): 6277 news averaging 71 words covering 464 stocks listed in the Standard & Poor 500 for the years 2000-2002. The automated labelling process is described in section 2.4. We have opted for a linear *Support Vector Machine (SVM)* (Joachims, 2001) approach as our classification algorithm with the software Weka[1]. All experiments have been cross-validated ten times.

2.1. Feature Types

We consider five types of features: *unigrams*, *stems*, *financial terms*, *health-metaphors* and *agent-metaphors*. The news are tokenized with the help of a POS tagger (Schmid, 1994). *Unigrams* consist of all nouns, verbs, adjectives and adverbs[2] that appears at least three times in the corpus. *Stems* are the unigrams which have been stripped of their morphological variants. The *financial terms* stem from a clinical study of investors discussion and sentiment (Das et al., 2005). The list comprises 420 words and their variants created by graduate students who read through messages[3] and selected words they felt were relevant for finance (not necessarily most frequent)[4]. *Health metaphors* are a list of words identified by (Knowles, 1996) in a six million word

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

[2] This list is augmented by the words *up*, *down*, *above* and *below* to follow (Koppel and Shtrimberg, 2004).

[3] The corpus was a random selection of texts from on Yahoo, Motley fool and other financial sites.

[4] Sanjiv Das, personal communication.

corpus from the *Financial Times* suggesting that the financial domain is pervaded by terms from the medical domain to describe market phenomena: examples include *addiction*, *chronic* and *recovery*. The full list comprises 123 such terms. Finally, recent work by (Morris et al., 2007) shows that in the case of market trends, investors tend to process *agent metaphors*, when the language treats the market as though it were an entity that produces an effect deliberately (e.g. *the NASDAQ climbed higher*), differently from object metaphors, where the language describe price movements as object trajectories, as events in which inanimate objects are buffeted by external physical forces (e.g. *the Dow fell through a resistance level*) or non-metaphorical expressions that describe price change as increase/decrease or as closing up/down (e.g. *the Dow today ended down almost 165 points*). The same study gives the verbs *jump*, *climb*, *recover* and *rally* as the most frequent indicators of uptrend movement, and *fall*, *tumbled*, *slip* and *struggle* as the most frequent indicators of downtrend movements. The point made in the study is that in the case of agent metaphors, investors tend to believe that the market will continue moving in the same direction, which is not the case for object metaphors or non metaphors. These results are potentially useful for sentiment analysis, as we are trying to find positively correlated textual features with market trends. To construct a list of potential agents, we extracted all nouns from our corpus and used WordNet[5] to filter out elements which were not hyponym of synset number 100005598[6], defined as *an entity that produces an effect or is responsible for events or results*: in this way we collected 553 potential agents. To allow those agents to carry out their actions, we completed this list with all 1538 verbs from the corpus.

2.2. Feature Selection

We consider three feature selection methods that (Yang and Pedersen, 1997) reported as providing excellent performance. Document Frequency (DF) is the number of documents in which a term occurs. We computed DF for each feature and eliminated features for which DF fell below a threshold (100). In Information Gain (IG), features are ranked according to a preferred sequence allowing the classifier to rapidly narrow down the set of classes to one single class. We computed the 100 features with the highest information gain. Finally, the χ^2 statistic measures the lack of independence between a feature and a set of classes. We computed the top 100 least independent features. It is worth mentioning that the same 100 features were selected using either IG or χ^2 statistic, except for a few features ranking order in the top ten.

2.3. Counting Methods

There are two methods worth considering for valuing each feature appearance in each news: the first is the binary method where a value of zero indicates the absence of the feature whereas a value of one indicates the presence of the feature. This method appears to yield good results in movie reviews (Pang et al., 2002). The second simply gives a count of the feature in the document and normalise the count for a fixed-length document of 1000 words (TF).

2.4. Classifying news using cotemporeneous prices: [-1 day,+1 day]

To construct our 500 positive examples we used similar criteria as (Koppel and Shtrimberg, 2004), based on contemporaneous price changes (stock price at opening the first market trading day after the news was published - stock price at closing the first trading day before the news was published):

- price change superior to the overall S&P index price change,
- price change in the interval [-4%,+4%] and
- price superior to \$10.

For instance, the following news about the company *Biogen, Inc.* (symbol BGEN), appeared on May 23rd 2002:

> Biogen, Inc. announced that the FDA's Dermatologic & Ophthalmic Drug Advisory Committee voted to recommend approval of AMEVIVE (alefacept) for the treatment of moderate-to-severe chronic plaque psoriasis.

At opening on the 24-May-2002, price reached \$48.43, whereas at closing on the 22-May-2002 it was \$38.71. Therefore, there is a positive price change of

$$\frac{\$48.43 - \$38.71}{\$38.71} = 0.19995$$

or almost 20%, the news is classified as being positive. The same reasoning is applied to find 500 negative examples, corresponding to a negative price change of at least 4%. The results of this experience is presented in table 1. The

Features		**F-Selection**		**F-Count**	
Unig.	67.5%	**IG**	67.5%	**Bin**	67.5%
Stems	66.9%	DF	59.4%	TF	67.6%
Fin.-T	59.2%	χ^2	66.1%		
Hea.-M	52.4%				
Age.-M	66.4%				

Table 1: Feature tuning

reference trio of parameters appears in table 1 between horizontal lines (unigrams, IG and binary). That is, in each successive measurement of accuracy, at least two values of the trio remained unchanged. For example, the classification accuracy when using stems, information gain and binary count is 66.9%. Strictly speaking, the best combination (unigrams, IG and TF) reached 67.6%, a tenth of one percent better than the basic trio (unigrams, IG and Bin). Given this non significant difference in accuracy and a favourable inclination for the binary method in the literature, we keep the basic trio as our parameter values for all other experiments. These results also show that features based on a list of agent metaphors describing market trend

[5] `http://wordnet.princeton.edu/`

[6] *causal agency#n#1*, *cause#n#4* and *causal agent#n#1*

movements appear more useful for the classification of financial news than a list of health metaphors or a human-constructed list of financial terms. At closer examination, it appears that most of the contribution is made by the notion of *agent*: only five of the eight most frequent indicators (*recover, climb, fall, slip* and *struggle*) actually appear in our corpus, and only one (*fall*) made the cut through the top 100 features that bring most information gain. We conjecture that the description of financial news retains the same agent-based feature as in market trend description, however it is expressed by commentators using a different set of (predicative) terms. In the remaining experiments we depart slightly from (Koppel and Shtrimberg, 2004) by taking into account negation, i.e. negated words (e.g. not rich) are featured as a single term (not_rich). We also remove all proper nouns as potential feature, our assumption being that a list of features without proper nouns is less tailored to a particular time-period, where some companies happen to be more in the spotlight than others.

Including Osgood's feature A study by (Mullen and Collier, 2004) have suggested that information from different sources can be used advantageously to support more traditional features. Typically, these features characterise the semantic orientation (SO) of a document as a whole (Osgood et al., 1957; Kamps et al., 2004). One such feature is the result of summing up the semantic relatedness (Rel) between all individual words (adjectives, verbs, nouns and adverbs) with a set of polarised positive (P) and negative (N) terms, for the domain of interest, here finance. This method can be expressed in the following formal manner:

$$\sum_{w}^{Words}(\sum_{p}^{P} Rel(w,p) - \sum_{n}^{N} Rel(w,n))$$

Note that the quantity of positive terms P must be equal to the quantity of negative terms N. To compute relatedness, we used the method described in (Banerjee and Pedersen, 2003) and WordNet[7]. The list of polarised terms we used follows:

```
Pos adjectives: good, rich
Neg adjectives: bad, poor
Pos nouns: goodness, richness
Neg nouns: badness, poverty
Pos verbs: increase, enrich
Neg verbs: decrease, impoverish
Pos adverbs: well, more
Neg adverbs: badly, less
```

Class	SO
0.00	-106
0.25	-89
0.50	-104
0.75	-114
1.00	-128

Table 2: Semantic Orientation

[7]Using the PERL package (Pedersen, 2004).

Although the relatedness measure is biased towards negative, as illustrated by all negative semantic orientations, even for positive classes, the trend observed and expected is that positive classes are less negative than positive classes in general (coefficient of correlation is +0.76). This is a result conforting the validity of the automatic labelling technique. However, our result shows no significant improvement on accuracy (69%) if we include semantic orientation as one of our features.

2.5. Horizon Effect

The next experiment looks at the lasting effect of a news on the stock price of a company. Using 300 positive examples and 300 negative examples with a $\pm 2\%$ price variation, we computed classification accuracies for non cotemporeneous, more precisely subsequent, price changes. Therefore, news were classified according to price changes from the opening the first open market day after the news to X number of days after the news. We consider the following values for X: 2, 3, 7, 14 and 28. Table 3 presents the results. Given that classification accuracies are slowly worsening as

Horizon	Accuracy
[+1,+2]	69.5%
[+1,+3]	68.8%
[+1,+7]	67.5%
[+1,+14]	68.0%
[+1,+28]	66.3%

Table 3: Horizon Effect

we move further away from the day the news first broke out (coefficient of correlation is -0.89), we conclude that some prices are getting back to, or even at the opposite of, their initial level (i.e. before the news broke out). Assuming that in the interval no other news interfered with the stock price, this result also reinforced the validity of the automatic labelling technique.

2.6. Polarity effect

This experiment looks at the effect on accuracy a change in the labelling distance between two classes produces. The intuition is that the more distant two classes are from each other, the easiest it is for the classifier to distinguish among them, which translates as a higher accuracy.

Class 1	Class 2	Dist.	Acc.	Aver.
0.00	0.25	0.25	62.8%	
0.25	0.50	0.25	64.6%	
0.50	0.75	0.25	57.6%	62.3%
0.75	1.00	0.25	64.1%	
0.00	0.50	0.50	68.0%	
0.25	0.75	0.50	61.8%	66.4%
0.50	1.00	0.50	69.3%	
0.00	0.75	0.75	70.3%	
0.25	1.00	0.75	73.1%	71.7%
0.00	1.00	1.00	69.8%	69.8%

Table 4: Polarity effect

Table 4 presents classification accuracy using five classes:

- 0.00: 400 negative news, price change $<$ -2%
- 0.25: 200 negative plus 200 neutral news
- 0.50: 400 neutral news
- 0.75: 200 neutral plus 200 positive news
- 1.00: 400 positive news, price change $>$ +2%

The five classes above generate four possible combinations of labelling distance: 0.25, 0.50, 0.75 and 1.00. As expected, there is a positive correlation between labelling distance and accuracy (coefficient of correlation +0.89). This reinforces the validity of the automatic labelling technique.

2.7. Range Effect

The range effect experiment explores how the size of the minimum price change for a news to be labelled either as positive or negative influences classification accuracy. The intuition is that the more positive and negative news are labelled according to a larger price change, the more accurate classification should be. Table 5 shows results using cotemporeneous price changes. The labelling method yields once

Range	Nb examples	2-class	3-class
$\pm$0.02	1000	67.8%	46.3%
$\pm$0.03	1000	67.1%	47.9%
$\pm$0.05	800	69.5%	46.8%
$\pm$0.06	600	74.0%	50.1%
$\pm$0.07	400	76.3%	50.1%
$\pm$0.10	200	75.0%	51.3%

Table 5: Range Effect

again expected results: for two classes (positive and negative), the more comfortable the price change margin gets, the more accurate classification is (coefficient of correlation is +0.86). However, accuracies appear to reach a plateau at around 6%, where classification accuracy improvements beyond 75% seems out of reach. The last column of table 5 reports accuracies for the case where news whose price change is falling between the range are labelled as *neutral*. Although accuracies are, as expected, lower than for two classes, they are significantly above chance (33%). The same positive correlation is also observed between the price change margin and accuracies (coefficient of correlation is +0.88). In the next experiment we examine more in depth the effect of adding a neutral class on precision.

2.8. Effect of adding a neutral class on non-cotemporaneous prices: [+1 day,+2 days]

In all but one of the experiments so far, we have considered classes with maximum polarity, i.e. with a neutral class separating them. On one hand this has simplified the task of the classifier since news to be categorised belonged to one of the positive or negative extremes. On the other hand, this state of affairs is somewhat remote from situations occurring in real life, when the impact of news can be limited. Moreover, the information about overall accuracy of classification is not the most sought after information for investors. Let's examine briefly more useful information for investors:

Positive Precision A news which is correctly recognised as positive is a very important source of information for the investor. The potential winning strategy now available is to buy or hold the stock for the corresponding range. Therefore, it is very important to built a classifier with high precision for the positive class, significantly above 50% to cover comfortably transaction costs.

Negative Precision A news which is correctly recognised as negative is also an important source of information for the investor. The potential saving strategy now available to the investor, given that he or she owns the stock, is to sell the stock before it depreciates. Therefore, it is important to built a classifier with high precision for the negative class, significantly above 50% to cover safely transaction costs.

Positive and Negative Recall Ideally, all positive and negative news should be recognised, but given the potential substantial losses that misrecognition (implying low positive/negative precision) would imply for investors, only a decent level of recall is needed for both.

Table 6 gives a first glimpse of the sort of positive (+precision) and negative (-precision) precision we can expect if we built a 3-class classifier. In order to get closer to real classification conditions, we remove the constraint that stock prices must be greater than \$10. Results show that

Range	Nb examples	-Precision	+Precision
$\pm$0.01	1000	77%	51%
$\pm$0.02	800	41%	53%
$\pm$0.03	400	69%	54%

Table 6: Effect of adding a neutral class on non-cotemporaneous prices

precision is either worryingly close to 50% (the positive case), or is very volatile and could swing precision level well below 50% on too many occasions. Clearly, this demonstrate that if we are to build a financial news classifier satisfying at least high precision for the positive news, we must abandon the approach using three classes.

2.9. Conflating two classes

In section 2.8. we underlined the importance of high precision for the classification of positive and negative news and concluded that a 3-class categoriser was unlikely to satisfy this requirement. In this section we conflate two of the three classes into one and examine the effect on precision and recall. Table 7 displays three classification measures for the case where the classes neutral and negative have been conflated to a single class. Table 8 displays three classification measures for the case where the classes neutral and positive have been conflated to a single class. We used a range of $\pm$0.02, a forward-looking horizon of [+1,+2] days with 800 training examples. It is difficult to evaluate precisely what the cost of trading represents, but there seems to be enough margin of manoeuvrer to overcome this impediment, especially in the case of the positive classifier (table 7).

Measure/Class	POS	NEG+NEU
Precision	0.857	0.671
Recall	0.555	0.908
Accuracy	0.7313	

Table 7: Positive against all others

Measure/Class	NEG	POS+NEU
Precision	0.652	0.805
Recall	0.870	0.535
Accuracy	0.7025	

Table 8: Negative against all others

2.10. Positive and Negative features

Closer examination of the features resulting from the selection process paints a different picture from the one presented in (Koppel and Shtrimberg, 2004). Recall that (Koppel and Shtrimberg, 2004) used all words that appeared at least sixty times in the corpus, eliminating function words with the exception of some relevant words. We kept only adjectives, common nouns, verbs, adverbs and four relevant words, *above*, *below*, *up* and *down*, that appear at least three times in the training corpus. In a nutshell, (Koppel and Shtrimberg, 2004) found that there were no markers for positive stories, which were characterised by the absence of negative markers. As a result, recall for positive stories were high but precision much lower. Our findings are that negative and positive features are approximately equally distributed (53 negatives and 47 positives) among the top 100 features with the highest information gain and that recall and precision for positive stories were respectively lower and higher. We define sentimental orientation (positive or negative) of each feature as the class in which the feature appears the most often. Table 9 shows the top ten positive features and table 10 shows the top ten negative features. The *Pos* column indicates the position of the

Pos	Feature	+b/-b	+tf/-tf	+n/-n
1	common	29/8	33/13	1318/390
2	shares	33/11	48/17	2014/640
3	cited	20/4	20/4	427/49
5	reason	18/4	18/4	411/69
8	direct	7/0	7/0	163/0
9	repurchase	15/3	26/3	818/115
10	authorised	17/4	18/5	596/177
11	drug	6/0	6/0	114/0
13	partially	6/0	6/0	89/0
14	uncertainty	6/0	6/0	102/0

Table 9: Positive Features

feature in the top 100 ranking resulting from the information gain screening. The +b/-b column displays the number of documents (examples) in which the feature appears at least once (+b for positive and -b for negative). The +tf/-tf column displays the number of times the feature appears in the entire set of documents (+tf for positive and -tf for negative), while the +n/-n column displays the same values normalised to a constant document length of 1000 words.

Pos	Feature	+b/-b	+tf/-tf	+n/-n
4	change	0/8	0/11	0/206
6	work	1/11	1/12	33/183
7	needs	0/7	0/7	0/139
12	material	0/6	0/6	0/128
15	pending	0/6	0/7	0/100
16	gas	8/23	13/34	229/571
19	cut	1/9	1/10	15/216
20	ongoing	1/9	2/14	14/201
25	e-mail	0/5	0/5	0/68
26	week	0/5	0/5	0/72

Table 10: Negative Features

For example, the feature *common* appears in 29 positive examples and 8 negative examples. It also appears 33 times in all positive examples and 13 times in all negative examples. Below is one highly positive news (+11% price change) and one highly negative news (-49% price change) with positive features inside square brackets and negative features inside braces. The following news about the company Equifax Inc. (symbol EFX) appeared on the 20th of September 2001. Its stock price jumped from $18.60 at opening on the 21st of September 2001 to $20.70 on the 24th of September 2001, for a price change of 11.29%:

> Equifax Inc. announced that it is repurchasing [shares] in the open market, pursuant to a previous [repurchase] authorisation. The [Company]'s board of directors had [authorised] a repurchase of up to $250 million of [common] stock in the open market in January 1999, of which approximately $94 million remains available for purchase.

The following news about the company Applied Materials, Inc. (symbol AMAT) appeared on the 15th of April 2002; its stock price plummeted from $53.59 at opening on the 16th of April 2002 to $27.47 on the 17th of April 2002, for a price change of -48.74%:

> Applied Materials, Inc. announced two newly granted U.S. Patents No. 6,326,307 and No. 6,362,109, the [Company]'s third and fourth patents covering the use of hexafluorobutadiene (C4F6) {gas} chemistry for critical dielectric etch applications. A high-performance etch process chemistry, C4F6 used in an Applied Materials etch system, enables the industry's move to the 100nm chip generation and beyond.

3. Discussion

The surprisingly encouraging results we have presented for a forward-looking investment strategy should not be viewed outside its specific experimental setup conditions. In what follows we highlight a number of points worth considering:

Lack of independent testing corpus Cross-validation is a method which can provide a solid evaluation of the overall accuracy of a classifying method. However, a

more accurate evaluation should involve an independent testing corpus, ideally covering a distant time-period to avoid overfitting or overtraining. Nevertheless, we have attempted to avoid these caveats by keeping a small number of features compared to the number of training examples and by avoiding the use of proper nouns as features.

Pool of features Our pool of features was selected among the entire training set, which includes the cross-validated sections. Although to a small degree, this may have caused a *data-snooping* bias, where features were selected among the testing examples. On the other hand, as can be observed in tables 9 and 10, the interpretation of positive and negative features is not straightforward, which suggests that portability among different domains and even time periods could be problematic.

Size of documents Clearly, the size of documents is crucial for classification. The corpus we used averaged just over 71 words, which in general should be long enough to collect enough statistics. Nevertheless, if we look at our top ten positive stories (those with the highest positive price change), we found that half of them contained no feature at all, whereas three out of our top ten negative examples were similarly deprived of features. Given that this situation is likely to worsen if we train and test on different domains and periods, this is a potential area where a default bias can be difficult to avoid (i.e. a document without features will systematically be classified in the same class). One solution would be to increase the number of features.

Trading costs If the minimum transaction level to overcome fixed and relative trading costs is high, this brings upon the investors a burden of risk which he or she may not be able or willing to bear. The classifier should be characterised clearly by its level of precision matched with an estimate of the trading costs that would guide the investor in its decision.

4. Conclusion and future work

We have revisited a method for classifying financial news using automatically labelled data. Our findings give a different picture of the set of features best suited for the task and a somewhat less pessimistic prognostic as to the validity of such an approach for forward-looking investment. We indicate a number of elements where extensive research should be carried on to test the approach within a practical and realistic framework. To this end, our next step is to use our system coupled with a virtual trading site[8] to monitor financial news to invest in companies. This should give us a better idea of the effect of the transaction costs as well as the portability of the features and model developed during our experiments.

5. References

S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, 2003, pp. 805–810.*

Sanjiv Das, Asis Martinez-Jerez, and Peter Tufano. 2005. e-information: A clinical study of investor discussion and sentiment. *Financial Management*, 34(5):103–137.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, CZ, June. ACL.

Thorsten Joachims. 2001. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers.

J. Kamps, M. Marx, R. Mokken, and M. de Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *In LREC 2004, volume IV, pages 1115–1118.*

Francis Knowles. 1996. Lexicigraphical aspects of health metaphors in financial texts. In *Proceedings Part II of Euralex 1996*, pages 789–796, Department of Swedish, Göteborg University.

Moshe Koppel and Itai Shtrimberg. 2004. Good news or bad news? let the market decide. In *AAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 86–88. Stanford University, March.

Gilad Mishne. 2007. *Applied text analytics for blogs*. Ph.D. thesis, University of Amsterdam.

Michael W. Morris, Oliver J. Sheldon, Daniel R. Ames, and Maia J. Young. 2007. Metaphors and the market: Consequences and preconditions of agent and object metaphors in stock market commentary. *Journal of Organizational Behavior and Human Decision Processes*, 102(2):174–192, March.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Empirical Methods in NLP*.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of the 2002 Conf. on Empirical Methods in Natural Language Processing*.

Ted Pedersen. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *In Appears in the Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004.*

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Int. Conference on New Methods in Language Processing*, Manchester, UK.

D. Lawrie P. Ogilvie D. Jensen V. Lavrenko, M. Schmill and J. Allan. 2000. Mining of concurrent text and time series. In *6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, August 2000, August.

T. Wilson and J. Wiebe. 2003. Annotating opinions in the world press. In *In SIGdial-03.*

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In Douglas H. Fisher, editor, *Proc. of ICML-97, 14th Int. Conf. on Machine Learning*, pages 412–420, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US.

[8] `http://vse.marketwatch.com/`

An Unsupervised Method to Extract Topic Expressions from Reviews on TV shows

Takeshi S. Kobayakawa[†1,†2], **Jin-Dong Kim**[†2], **Jun'ichi Tsujii**[†2,†3]

[†1] Human & Information Science, NHK Science & Technical Research Labs.
Kinuta 1-10-11, Setagaya-ku, Tokyo 157-8510, Japan
[†2] Department of Computer Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan
[†3] School of Informatics, University of Manchester
POBox 88, Sackville St, MANCHESTER M60 1QD, UK
kobayakawa.t-ko@nhk.or.jp, jdkim@is.s.u-tokyo.ac.jp, tsujii@is.s.u-tokyo.ac.jp

Abstract

This paper presents an unsupervised method to extract topic expressions from TV show reviews. We propose an unsupervised approach, which does not require any elaborated linguistic resources, *e.g.* annotated corpora, syntactic parsing, etc. We also present a manually-annotated corpus used as benchmark data, in order to evaluate the proposed method. Experimental results reveal that the proposed simple method can be effectively used to winnow topic expressions from a large number of reviews.

1. Introduction

The widely spreading internet technology has enabled individual users to actively express their opinions to the public. It has led to an abundance of opinions in the public space, the Internet. To companies, it is becoming more and more important to collect and analyze the opinions on their products to better respond to the demand from their potential customers.

When reviews for content-based products such as books, movies, or TV shows are analyzed, the topic of the sentiment tends to be expressed as phrases or clauses rather than as simple nominal expressions that denote product names or specific properties. Thus, topic identification becomes a complex but necessary step. As the study(Kim and Hovy, 2006) states, opinion topics can refer to social issues, government's acts, new events, or someone's opinions for online news media texts. For example, in our domain, *I was very impressed by Dr. Nakamura's statement, "Think from the viewpoint of other people,"* is a review sentence including a noun clause (*Dr. Nakamura's statement "—."*) as an opinion topic.

There are a growing number of responses by several channels, including direct responses in e-mail or web form, or indirect responses in reviews on web pages or on (we)blogs. Autonomously produced feedback are indirect responses, and are thus indirectly delivered to the company. Since they are buried in a large amount of heterogeneous web contents, it is necessary to identify the opinion-holding sentences. On the contrary, direct responses are opinions that are delivered directly to the company. A well-designed question form can produce a controlled feedback, making most of the sentences as opinions. In this situation, the problem falls into an easier one, because extracting opinions is not necessary any more.

In this study, we analyze the TV show reviews in Japanese. The task is to analyze the reviews that are directly collected, and clearly identify what part of the TV show the opinions are about. As a characteristic of the task, extracting opinion-holding sentences and identifying TV shows according to opinions, is unnecessary, while topic identification is still necessary.

For example, even in consideration of one TV show, some opinions can be about the characters, while others can be about the facts described by the program. As for the characters, some opinions can be about specific quotations, while others can be about the attributes (such as the sincerity) of the character. The opinion topic, which has a reference to the TV show, varies widely from opinion to opinion.

Finally, we break up the task into two parts:

- topic expression identification
- sentiment classification

We focus on identifying the topic expression of the opinion in this study. It is a complement to the previous work done on sentiment classification (Kobayakawa et al., 2007).

We propose a method that would make use of the characteristics of the task — There are many reviews to

one specific TV show, and there are also several TV shows for which reviews are given. We assume that a review is composed of topic expressions and sentiment expressions, and that the distributions of sentiment expressions and topic expressions are different over review documents: sentiment expressions will be rather evenly distributed while topic expressions will "burst" in each review according to the content being reviewed. Based on the assumption, we propose a method to extract topic expressions, incorporating a statistical hypothesis test and corpora comparison.
Experimental results using TV show reviews written in Japanese show the proposed method is effective in winnowing topic expressions from opinion-holding sentences. Although the experiments were performed with Japanese texts, the proposed method should be applicable to any other language, since it does not involve any language-specific processing.
This paper is organized in the following way: Related work is discussed in section 2.. The corpora that we built and topic expressions are described in section 3.. The method of extracting topic expressions is explained in section 4., the experiments conducted are in section 5., discussion is in section 6., and our conclusions and future work are in section 7..

2. Related work

Sentiment analysis has been the focus of attention in recent years, which is a combination of several technologies, such as subjectivity detection, opinion holder identification, review classification, and topic identification. Subjectivity detection is used to decide whether a given sentence has an opinion (Wiebe et al., 1999). Opinion holder identification is used to decide who said the opinion (Bethard et al., 2004). Review classification classifies the kind of the opinion (Turney, 2002). Topic identification identifies what the opinion is about (Kim and Hovy, 2006).
When sentiment analysis is applied to different domains, different methods should be adopted for the technologies, depending on what domain the sentiment analysis is applied to. One example is a typical sentiment analysis for product reviews (Kobayashi et al., 2005). Topic identification chooses one noun from the pre-defined series of products or their attributes, and review classification decides the polarity of the opinion. Another example is an analysis for online news media texts (Kim and Hovy, 2006). Review classification again decides the polarity of the opinion. However, opinion holder identification and topic identification classifies the result of semantic role labeling. Little research has been conducted for opinion topic identification, except the study(Kim and Hovy, 2006). In the study, opinion topics are identified on top of semantic role labeling. In English, an analysis based on Frame Semantics like FrameNet (Baker and Sato, 2003) is widely available. Thus, a method based on the semantic role labeling is a realistic choice. However, in other languages like Japanese, semantic role labeling is in itself, a elaborate task. Although previous works(Kawahara and Kurohashi, 2006) exist in this area, the study concentrates on building the case frame dictionary, and not on semantic role labeling. What we really need is semantic role labeling, which is still missing. We therefore adopt an unsupervised approach without using semantic role labeling.
The unsupervised approach makes use of the characteristic of the task. Since the sentences are collected by channel only reviews come, all the sentences that come are reviews for a specific TV program. Opinion holder is the sender of the e-mail or web forms, and is anonymous or already identified. Although the name which the TV program reviews are about is known, the topic of the opinion is unclear. Without any deep analysis, frequently used expressions are extracted. Then, detecting the topic of the opinion is achieved by comparing the extracted expression from corpus to corpus. The proposed method compares the corpora between focused corpus and other corpora. General discussion for comparing corpora can be found in (Kilgarriff, 2001).

3. Corpora of TV show reviews

We built corpora of reviews on TV programs using pseudo responses. We asked a group of people to watch four TV programs, as shown in Table 1, and make comments about them. These programs are a series of cultural programs concerning different topics. The number of people is between 120 to 130, depending on the program, and each of them was asked to provide 5 to 10 opinionated sentences for each program. While the expressions that specific viewers tended to use may differ from one viewer to another, using a high number of people reduces the inconsistencies in the expressions that the viewers tend to use, and thereby normalizes the data. The number of sentences and words of reviews are shown in Table 2[1].

We use this corpora to evaluate our methods to extract topic expressions. We mainly focus on the viewers'

[1]All the data is in Japanese. Because Japanese is not written with spaces between words, the sentences need to be segmented into words by a morphological analyzer. We used `chasen 2.3.3` and `ipadic 2.6.3` from `http://chasen.naist.jp`

ID	*Contents*
$\mathcal{A}$	Volunteer activity by a Japanese doctor in Afghanistan
$\mathcal{B}$	Account books of Samurai
$\mathcal{C}$	Club baseball team run by a Japanese comedian
$\mathcal{D}$	Biography of Mozart as told by a Rakugo storyteller

Table 1: 4 TV programs that were watched

ID	# sentences	# words
$\mathcal{A}$	916	19,396
$\mathcal{B}$	775	14,546
$\mathcal{C}$	762	14,461
$\mathcal{D}$	770	13,534

Table 2: The properties of the corpora of reviews on TV programs

comments on program $\mathcal{A}$ in Table 1, "Voluntary activity by a Japanese doctor in Afghanistan," [2] while three other programs were used as a reference described in subsection 4.2.2..

Topic expressions are the parts in the opinion where the object of the opinion is mentioned as referencing the contents of TV programs. The corpora are annotated[3] where the reference to the TV program contents are made. Several sample sentences are given here with the topic expressions indicated by the underline:

Example. *I was very impressed by Dr. Nakamura's statement "Think from the viewpoint of other people."*

Example. *I was inspired by the quote "A person is worth loving."*

Example. *I realized the importance of water because it makes poor-looking soil turn green.*

The properties of the topic expressions of a corpus are shown in Table 3 in the average scores. The number of topic expressions per sentence can be more than one, so the average is 1.23. Furthermore, the topic expressions averagely consists of nearly six morphemes, which shows that they are more than simple nouns but chunks of expressions.

To observe the details of the complexity of the topic

# topic expressions / sentence	1.23
# words / topic expression	5.94

Table 3: The properties of the topic expressions of a corpus

expressions, we created a hierarchical thesaurus based on (Ikehara et al., 1999) as in Figure 1, and added up the occurrences of the elements of the thesaurus. The reason we created the thesaurus, was to make the categories the same for both nominized expressions and for (non-nominized) verbal expressions; we wanted to avoid top-level branches of nouns and verbs that were found in the original thesaurus(Ikehara et al., 1999).

Every topic expressions was extracted and assigned to one of the leaves in the thesaurus by hand, word for word. The breakdown of the components is shown in Figure 2. The elements below *Matter* tend to compose phrases or clauses, and comprise nearly 42% of the parts. The statistics support our approach to extract clauses or phrases as a whole rather than as individual words.

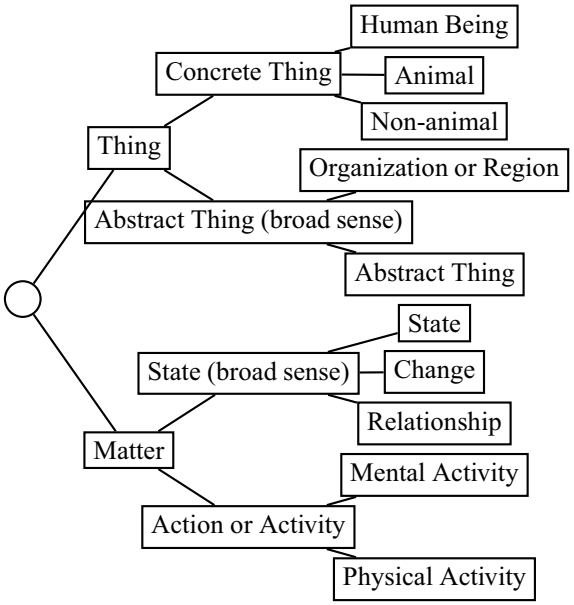

Figure 1: Thesaurus for classifying topic expressions

4. Methods for extracting topic expressions

We propose two types of methods and their combination to extract topic expressions. The procedure of the proposed methods are shown in Figure 3.

4.1. Word-based extractions

Term frequency-inverse document frequency (tfidf) were used to extract topic specific keywords. The extractions were based on words, so the amount of word occurrences were sufficient to calculate those statistics.

[2]The program was broad-casted on July 24/2006, entitled "Satisfaction from knowledge." Dr. Nakamura, a Japanese physician helped locals dig a well to obtain water as part of volunteer work in a region of Afghanistan that had been devastated during the war. He also ran a clinic there. Dr. Nakamura says in an interview, "Think from the viewpoint of other people," "A person is worth loving," and "Sincerity is worth believing in." These quotes were partially extracted in Table 5.

[3]One person was asked to annotate the content-referencing parts subjectively after watching the TV program. The annotated corpus was the test set for the experiments, and we evaluated how many of them were correctly identified.

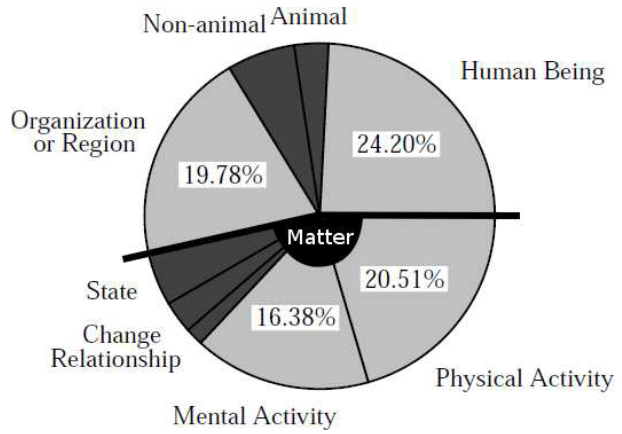

Figure 2: Breakdown of topic expressions

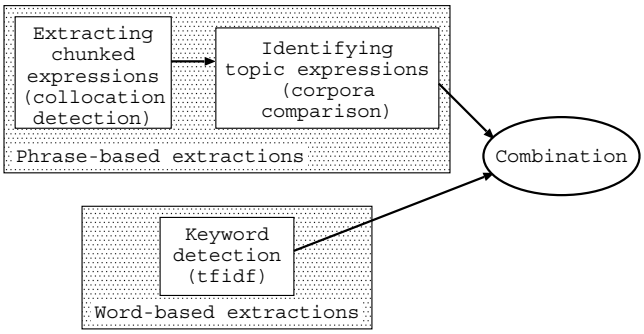

Figure 3: Procedure for extracting topic expressions

4.2. Phrase-based extractions

The extractions based on phrases were also done. Because the phrase occurrences are not sufficient for statistics, we used a different approach.

We first assume that there are two parts in TV show reviews; The first types are expressions by which viewers express their opinions. These expressions are not dependent on the parts of TV programs for which comments are made. The second types are topic expressions, to which viewers express their opinions. We herein propose a method based on techniques for detecting collocations(Manning and Schütze, 1999) and comparing corpora (Kilgarriff, 2001). First, we extracted chunked expressions without parsing sentences. Then, we examined whether the expressions are general or specific on that issue.

4.2.1. Extracting chunked expressions

We chose a hypothesis testing method based on t-statistics (Manning and Schütze, 1999) to extract chunked expressions. The method is basically for detecting collocations.

Words w^1 and w^2 have occurrence probabilities $P(w^1)$, and $P(w^2)$, independently. They also have a co-occurrence probability $P(w^1w^2)$, which is the probability that bigram w^1 is followed by w^2. The null hypothesis is that the two words do not form a collocation — that is, that the bigram appears by chance;

$$P(w^1w^2) = P(w^1)P(w^2). \quad (1)$$

We performed a statistical hypothesis testing regarding the probability of a certain constellation occurring. The t statistic is calculated as:

$$t = \frac{\bar{\chi} - \mu}{\sqrt{\frac{s^2}{N}}}, \quad (2)$$

where $\bar{\chi}$ is the sample mean, s^2 is the sample variance, N is the sample size, and μ is the mean of the distribution. In this case, the $\bar{\chi}$ is the bigram probability $p(w^1w^2)$, and the μ is the product of the unigram probabilities, $p(w^1)$, and $p(w^2)$. For a multinomial distribution, the variance s^2 is approximated by small probability p, as

$$s^2 = p(1-p) \approx p. \quad (3)$$

The t statistic is easily extended to calculate an arbitrary length n-gram:

$$t = \frac{\bar{\chi} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{p_n(w^1 \cdots w^n) - p_1(w^1)p_1(w^n)}{\sqrt{\frac{p_n(w^1 \cdots w^n)}{N}}} \quad (4)$$

Using the algorithm in Figure 4, we extracted chunked expressions for up to 10-grams. The loop started with 10-grams to detect collocations, decreasing its n-gram length. If the detected expression was a part of a longer expression already detected, then the shorter one was not accepted.

- for n = 10 to 2
  - Choose one n-gram from w_n
    - Calculate occurrence probability p_n for w_n
    - Decompose w_n to 1-gram and calculate its product of probabilities p_1
    - Use eq.(4) to calculate t from p_1 and p_n
    - If t exceeds significance level, accept w_n as a candidate
    - For all N such that $N > n$, if w_n is a sub-string of w_N, exclude w_n from the candidate
    - If w_n remains the candidate, print it
  - Continue loop on w_n
- Continue loop on n

Figure 4: The algorithm for extracting chunked expressions.

4.2.2. Identifying topic expressions

What was extracted with the method described in subsection 4.2.1. was compared to what was extracted from reviews of other programs. Each expression extracted from reviews of non-focused programs was subtracted from the expressions extracted from reviews on the focused program. Because an exact match of the phrases was too strict, we loosened the conditions somewhat when subtracting the expressions; for two n-grams and for a certain l, if subsequence l-grams of the two n-grams matched for any part, then the two n-grams were treated as a match.

4.3. Combination of word-based extractions and phrase-based extractions

The word-based extractions and phrase-based extractions were combined. Topic specific expressions can also be detected by a conventional term frequency-inverse document frequency (tfidf) based method, which is widely used in information extractions. A threshold was used for the tfidf value, among which the expressions were considered to be topic specific. The unit of tfidf-based extraction is words. When the method described in subsection 4.2. were combined with the method in subsection 4.1., we simply added the set of extracted expressions from both methods.

5. Experiments

Experiments were conducted using the method described in section 4. using the manually annotated corpus, which is described in section 3., as a benchmark data.

This section describes a comparison of word-based extractions, phrase-based extractions, and a combination of them. The baseline is the extreme case if all of the expressions are identified as the topic expressions.

5.1. Experimental Setups

For word-based extractions, the term frequencies were calculated with the reviews on the program $\mathcal{A}$, and the document frequencies were calculated with the reviews on the programs $\mathcal{B}, \mathcal{C}$, and $\mathcal{D}$.

For phrase-based extractions, the significance level for the t test was set to 0.5%. The expressions were extracted by the method described in subsection 4.2.1.. The number of extracted expressions for $\mathcal{A} - \mathcal{D}$ are shown in Table 4. Some of the intermediately extracted chunked expressions as described in subsection 4.2.1. from reviews on program $\mathcal{A}$ are shown in Table 5. Then, the extracted expressions were filtered by the method described in subsection 4.2.2.. The reviews on the programs $\mathcal{B}, \mathcal{C}, \mathcal{D}$ were used to subtract common expressions. The number of remaining expressions differs depending on the length, l, described in subsection 4.2.2.. If l is three, 70 expressions remain, and if l is four, 25 expressions remain. We chose l to be three, because of the deterioration in the performance for other values. The remaining expressions were identified as a topic expression, finally extracted by phrase-based extraction. They are indicated by the gray background in Table 5 and some more of them alone are shown in Table 6.

ID	# extracted expressions
$\mathcal{A}$	234
$\mathcal{B}$	170
$\mathcal{C}$	166
$\mathcal{D}$	161

Table 4: The number of extracted expressions

Occurrences	*t score*	*Expressions*
7	2.65	"Think from the viewpoint of other people
10	3.16	I was impressed by
16	4.00	I thought that
12	3.46	with human and human
11	3.32	I came to think
11	3.32	I was impressed by
10	3.16	Isn't it that
8	2.83	I imagine that
7	2.65	is worth loving
7	2.65	Is it doing that
17	4.12	I felt
16	4.00	doing
11	3.32	the importance of water
11	3.32	I had a feeling that
11	3.32	for Japanese young people

Table 5: Intermediately extracted chunked expressions as in subsection 4.2.1. from reviews on program $\mathcal{A}$. The gray background indicates what was finally extracted by our phrase-based extraction, while the underline indicates what was extracted by word-based extraction. The expressions *other people* and *young people* are one word in Japanese.

5.2. Results

The number of topic expressions that were correctly identified was evaluated. The precision, recall and F-

Expressions
"Think from the viewpoint of other people
with human and human
is worth loving
the importance of water
to Japanese young people
the importance of
of NGO

Table 6: Identified expressions for reviews on program $\mathcal{A}$.

measure were defined as

$$\text{precision} = \frac{A}{A+C}, \qquad \text{recall} = \frac{A}{A+B} \tag{5}$$

$$\text{F-measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{6}$$

where A is the number of correctly identified topic expressions, B is the number that could not be identified, and C is the number that was misidentified. The results with the units of words and with the units of characters are shown in Table 7. The combination is

	units	word	character
Baseline	precision	33.41%	36.69%
	recall	100%	100%
	F-measure	50.09%	53.68%
Word-based extraction	precision	55.78%	52.78%
	recall	38.26%	53.67%
	F-measure	45.39%	53.22%
Phrase-based extraction	precision	59.34%	55.18%
	recall	15.55%	16.93%
	F-measure	24.65%	25.91%
Combination	precision	52.90%	51.66%
	recall	48.25%	60.48%
	F-measure	50.47%	55.73%

Table 7: The results on how many expressions were correctly identified. The baseline is the extreme case if all of the expressions are identified as the topic expressions. The figures of precisions and recalls are where F-measures are the best.

the best performance in F-measure. When the combination is compared to the baseline, precisions are improved in the cost of recalls. If some number of samples of topic expressions are required to be extracted from the corpus, the combination can help reduce the number of samples of reviews.

The precision versus recall curve is shown in Figure 5, when the threshold for tfidf was varied. The word-based extractions are indicated by a solid line, while the combination of word-based extractions and phrase-based extractions is indicated by a dashed line. The combined extractions showed a better performance.

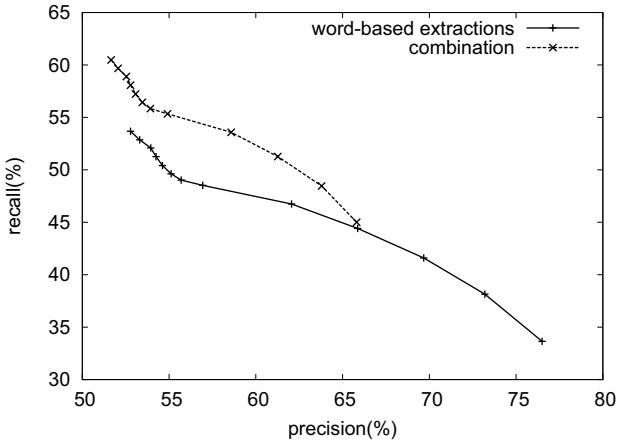

Figure 5: Precision-recall curve for word-based extractions, and their combination with phrase-based extractions (word units)

The curves of F-measures are shown in Figure 6. The horizontal axis shows the number of terms above the tfidf threshold; the maximum F-measures are shown in Table 8. Our method combined with the tfidf-based method outperforms the tfidf-based method alone by 2.6% to 5.0%.

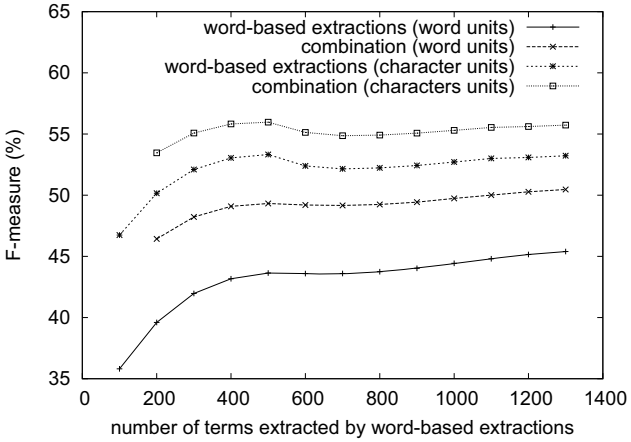

Figure 6: Curve of F-measures vs. the number of terms extracted by word-based extractions.

unit	words	characters
Word-based extractions	45.39%	53.32%
Combination	50.47%	55.97%

Table 8: Maximum F-measures

6. Discussion

Although the F-measure do not show significant difference between the experimented approaches including baseline one, their performance is very different in terms of precision and recall. For example, in terms of recall, the baseline approach will perform the best, of course, since in this approach the whole reviews are retrieved to be scanned. Sometimes, sampling of the reviews are effective, where precision should be taken into consideration. In terms of precision, the phrase-based extraction will perform the best. Thus, it can be a good choice when one wants to quickly sample topic expressions regardless of the recall. The word-based extraction and its combination with phrase-based extraction balance between the two extremes. Compared to the word-based extraction, the phrase-based extraction and its combination with word-based extraction produce results of higher readability.

Phrase-based extractions can detect the expressions as chunks. Thus, combining them with the word-based extractions improved the performance. Most of the improvements occurred when functional words cannot be detected by the word-based extractions. However, the improvements are not restricted to functional words; For example,

word-based extractions with the people in the field

phrase-based extractions with the people in the field

the word-based extractions only extract the expression, *the field* [4], while phrase-based extractions extract the expressions, *the people in the field*. Not only functional words like *in*, but also non functional words like *the people* are correctly extracted. As seen in Table 5, word-based extraction only extracts *water* while phrase-based extraction extracts *the importance of water* as a whole, for example. This improves the readability of the topic expressions, and leads to an easy guess of the opinion topic.

The amount of the text extracted is very different depending on the method. Compared to the baseline, the number of word count extracted by phrase-based expression is 1.2%, while those by the combination method is 3.7% at its best. This means, if the combination method is adopted as a system, the necessary amount of the text to be read can be reduced to 3.7% for similar F-measure performance as the baseline.

The method used in this study does not use a language specific analysis, so it is considered to be language neutral. As an experiment, only reviews in Japanese are analyzed; however, the method is expected to be easily applied to the other languages.

Our phrase-based extractions only extract expressions that are more frequent than thresholds. So, less frequent expressions are unlikely to be extracted. To extract them, we need to analyze the structure of the sentence, perhaps by parsing the sentence, or by template-based matching of the sentence, or by labeling the semantic role(Kim and Hovy, 2006). Although it is not comparable to the study(Kim and Hovy, 2006) since the test set is different, the proposed method in this study seems competitive to the study for the same task.

7. Conclusions and future work

We first built corpora for evaluating reviews for TV shows. In the corpora, sentiment annotations are conducted from the viewpoint of 1) opinion type classification, and 2) opinion topic identification. Opinion topics are specified as topic expressions, and are analyzed using a simplified thesaurus to make clear how much of the topic expressions are beyond simple nouns.

Then, we described a method for analyzing reviews on TV shows. We proposed a method that makes use of the characteristics of the task. The method was based on techniques of tfidf, collocation detections, and corpora comparisons. Phrase-based extractions, which use collocation detection techniques and corpora comparisons, can extract expressions as chunks. As a result, many topic expressions could be extracted as a whole. Since the phrase-based extractions are based on statistical hypothesis testing, only the topic expressions mentioned by many reviewers are extracted, causing a low recall. On the other hand, conventional word-based extractions performed with a better recall. Finally, the combination of word-based extractions and phrase-based extractions outperformed other methods. Experimental results suggest that extracted expressions are reasonably identified as the opinion topic.

Apart from the performance of topic extraction accuracy, the proposed method had another effect. The combination could reduce the necessary amount of the text to be read to 3.7%.

The experimental results show the proposed method is effective in winnowing the topic expressions from opinion-holding texts. Since it does not involve any language-specific processing, there is a high chance that the proposed method can be applied to other languages. Our future work will include seeking any synergic effect when the proposed method is combined with other linguistic analysis-based methods, *i.e.* use of syntactic parsers, semantic role labeling, etc.

[4] No determiner exists in Japanese, so *the field* is one word, and it can be extracted with the word-based extractions.

8. References

Collin F. Baker and Hiroaki Sato. 2003. The framenet data and software. In Poster and Demonstration at Association for Computational Linguistics.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileis Hatzivassiloglou, and Dan Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1999. Goi-taikei — a japanese lexicon cdrom.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In In Proceedings of the 5th International Conference on Language Resources and Evaluation.

Adam Kilgarriff. 2001. Comparing corpora. International Journal of Corpus Linguistics, 6(1):1–37.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In Proceedings of the Workshop on Sentiment and Subjectivity in Text, pages 1–8.

Takeshi S. Kobayakawa, Masaru Miyazaki, Mahito Fujii, and Nobuyuki Yagi. 2007. Detecting sentimental expressions with part-of-speech n-grams (in japanese). Proceedings of The Thirteenth Annual Meeting of The Association for Natural Language Processing.

Nozomi Kobayashi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2005. Opinion extraction using a learning-based anaphora resolution technique. The Second International Joint Conference on Natural Language Processing, Companion Volume to the Proceeding of Conference including Posters/Demos and Tutorial Abstracts:175–180.

Chiristopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. The MIT Press.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 417–424.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In Proceedings of ACL-99, pages 246–253.

Multilingual Metaphors

POSTER

Politics Makes the Swedish :-) and the Italians :-(

Jerom F. Janssen & Carl Vogel

Computational Linguistics Group
Intelligent Systems Laboratory
O'Reilly Institute
Trinity College Dublin
Dublin 2, Ireland
jfjanssen@gmail.com & vogel@tcd.ie

Abstract

This paper explores the extent to which intercultural differences in verbal and nonverbal feedback are manifest in text-only communications. Emoticons are the broad label for nonverbal cues often provided by writers of informal emails and newsgroup postings. Cross-cultural differences in interpersonal communication have long been reported for face-to-face contact. This paper explores the transference of those strategies to the visual-only domain of email text. We focus on a sampling of newsgroup discussions of politics and science in Swedish, German, Italian and English. Analysis of other communication channels suggests that extremes of behavior should occur for Swedish and Italian. We hypothesized on the basis of that research that Swedish news postings should be well adjusted to provision of emoticons in the absence of true visual feedback of body language, and that Italian postings should display least adjustment to the loss of an information channel. The results were surprising in that the number of positive, negative and neutral emoticons relativized to the total number of postings was not significantly different for Swedish and English. Interesting divergences appeared for German (twice the positive emoticons as Swedish) and Italian, nearly (twice the negative emoticons as in Swedish). These and other results are reported in more detail. The results can be interpreted as demonstrating a direct transfer of intercultural differences into informal electronic communications.

KEYWORDS: emoticons, inter-cultural differences, sentiment analysis

1. Background

Allwood (1985) discusses a range of issues that interact in inter-cultural communication that can at times lead to communication breakdown. Interesting about noticing differences in communication strategies, inter-culturally, is that it points out within-culture patterns of communication and creates questions about whether those patterns persist across all communication channels, all other things held equal. Among the many axes of variation, Allwood (1985) has drawn attention to feedback mechanisms, whether supplied for auditory or visual channels during interpersonal communication. He notes that in Japanese culture, direct eye-contact is avoided in a politeness strategy, and that auditory feedback is much more the norm, even when interlocutors are together in person. Allwood (1999) notes that in Swedish, too, auditory feedback is more pervasive than visual facial feedback.

For a contrast in the other direction, it is common to remark on the status of visual feedback in Italian (cf. Albrecht et al. (2002)); one expects an ample supply. However, the two forms of feedback are not necessarily inversely proportional to each other. Cerrato (2003) has compared Swedish and Italian spoken dialogues for auditory feedback and notes a vast difference in the propensity for feedback to overlap with a primary contribution of a dialog partner in the Italian part of the data. This is compatible with there being longer pauses between utterances in the Swedish data, in accounting for that difference; however, because the analysis reported is only of transcripts, there is no evidence that the increased auditory feedback is not also accompanied by visual feedback as well.

Pretheoretical ideas (that is, stereotypes) of communication in various languages leads one to wonder about how those stereotypes might transfer to alternative communication modes in which one form or other of the communication is unavailable. Thus, we here explore a topic at the intersection of sentiment analysis and inter-cultural studies. We have decided to examine informal electronic communication as is constituted by newsgroup interactions. We have focussed on text-only asynchronous communications (excluding email, blogs and multi-modal postings). We explore whether in that category significant differences emerge in the use of emoticons among within-linguistic community communicators using Swedish, German, Italian and English.[1] Sometimes, for convenience of expression, we will talk as if all users posting to, for example, the .de subnet actually are German, even though it is clear that this is not the case. A larger caveat is that while we are referring to emoticons as if they are feedback mechanisms simpliciter. They can, of course, be used asynchronously to comment on a bit of text supplied by one's correspondent, but they are also used to indicate how one's own sentence should be read (perhaps with irony, perhaps as an indication of humor, etc.) as can facial expressions accompanying utterances. Strictly, these are not instances of feedback on the words of another, but an interpretation guide for one's own statements.[2] We are simplifying enormously to anticipate that what happens during in-person communication (lack of visual feedback leads to increased auditory feedback) will have a manifest analogue in a domain that lacks

[1] To control the data as much as possible, the English sampling was from the .uk newsgroup hierarchy. More on these details will emerge in §2., which discusses our methodology.

[2] A difference between emoticons and visually provided interpretation guides is that communicators have more conscious control over emoticons than their body language.

in-person visual and auditory feedback altogether; the written communication permitted by newsgroups is lacking in exactly this respect, although emoticons may prove to play a part in the analogy.

Based on the original observations of Allwood, we anticipated that differences in deployment among those linguistic communities would obtain. However, we did not have a strong *a priori* intuition about the direction of effect. To the extent that a hypothesis was articulated, it was that one of the authors extrapolated from Swedes' using more verbal than visual communication in providing feedback during communication in direct contact: when using a communication channel devoid of facial or aural monitoring possibilities, they could be more accustomed from the outset to using verbal cues (even if reading is inherently visual) and thus perhaps inclined to provide textual renderings of visual cues as emoticons,[3] and perhaps relatively more so than Italians would. That is, one possibility is that the Swedes would be accustomed to something like emoticons already, and that the Italians would be frustrated by the constraining medium for informal discourse. However, it seemed equally likely that things could go in a different direction. The results tend towards the latter prediction.

The divergence from expectations in the electronic medium may have been easy enough to predict from the fact that "flame wars" are fairly uniquely confined to email and news groups, and that people are known to engage in those verbal battles with such vitriol that one may find it difficult to believe that the author is the same person one shares coffee breaks with. However, while an e-Swede may be different from a Swede in terms of the verbal feedback supplied, the question explored here is whether e-Swedes differ from Swedes in as predictable a way as do e-Italians from Italians, and so on.

We selected Swedish, Italian, English and German as languages through which to explore the actual state of affairs in informal interaction through newsgroups. This asynchronous medium is less interactive than internet chat might be, but it can be at least as interactive as email. We describe the method for selecting newsgroups in more detail in §2. In the end, the results we present control for those four languages, and two different sorts of topics (politics and science). In §3. we detail the data collection and filtering process in greater length. Of particular interest here is the selection of emoticons to be considered, as discussed in §3.2.3. We present the results and discuss their potential implications in §4. Finally, we suggest what looks most promising as a way to proceed in inter-cultural sentiment classification using emoticons on the basis of this work, proposing improved interactivity metrics in §5.

2. Method

We compared the usage of emoticons in terms of relative frequency in eight newsgroups: for four languages—English, German, Italian and Swedish—we compared two newsgroup topics—politics and science—which we think are suitably contrastive in nature. The languages were chosen to be plausibly at the extremes of communication using visual feedback during in-person communication, with English and German as intermediate control languages. Given the dominance of English as a language for newsgroup communications, we chose English as a baseline, thinking that trends in other languages would be at least partially influenced by the trends there. We selected newsgroups whose network hierarchy structure revealed their language focus (German: .de; Italian: .it; Swedish: .se, .swnet; English: .uk). Initially we thought we would be able to survey each of these languages using a broader range of subject categories spanning Politics, Culture, Science, Computers and Pets. However, in the end, only Science and Politics had clearly representative counterparts in each language's news hierarchy. We recorded the source within that hierarchy of each news posting, but we did not analyze data by newsgroup at any finer level of granularity than the topic areas just named, five, reduced to two.

A set of frequently used emoticons was decided upon as outlined in §3.2.3. We wanted to work with the emoticons in terms of a three-way classification of them as expressing positive feedback, negative feedback, or neutral feedback, irrespective of context. The basic method, then, is to examine the relative frequency of the various sorts of emoticons as a function of language and subject category to spot whether there is a significant difference when taking into account the size and interactivity of the respective newsgroups. That is, we obviously did not want to compare raw frequency. However, we additionally wanted to know if the use of emoticons was in any way influenced by degree of interactivity in the postings. Interactivity can be operationalized in any number of ways: the ratio of distinct posters in a newsgroup (which correlates with its outreach) to the total number of postings; the average number of actual newsgroup cross postings per message; the average number of included messages in postings (this latter figure has to be relativized to the total number of postings in the area as there cannot be replies to more messages than there were in the first place). The final figure is also indicative of the depth of readership of a newsgroup in that it correlates with the number of people actually reading each other's posts and taking discussion into an involved exchange of ideas. These metrics are discussed further in §3.3. A few more interactivity metrics are discussed in §5. It is not possible in the context of this paper to report on all of the statistics that are possible to extract from the dataset. The methods here are clearly oriented towards statistical comparison of frequency distributions rather than detailed functional analysis (cf. Allwood and Cerrato (2003)).[4]

3. Data

3.1. Collection & Preparation

We used two kinds of data: a collection of emoticons (discussed in §3.2.3.), and a collection of Usenet newsgroup-postings. The postings were taken from a news-server that

[3] Note that the verbal feedback is partial "hmmms" & such, and not fully verbal reconstructions articulating the content of feedback.

[4] In §3.2.3., we comment on the problem of false positives in counting emoticons. At the present time this is a source of noise that we hope at best to estimate, and perhaps in future work, find a compelling method to eliminate.

has been archiving Usenet posts (excluding binaries) since late September 2006 (the HEANET in Ireland). As a result, we had access to a plethora of topics, but the posts represent a relatively limited time span. The dates of our final selection range from September 2006 up to February 2008.

3.2. Treatment of the data

There has been no manipulation during the pre-processing stage of the Usenet data, except for the automatic exclusion of messages deemed to be spam by SpamAssassin.[5] Within the remaining hierarchy, postings to subgroups were collected into a single directory for the topic, recording the original group submission.

3.2.1. Parsing the data

Newsgroup messages consist of two parts, a Header and a Body. The header contains meta information only, such as who posted the message, to which group(s) was the message posted at what date and time, did the message start a new thread or was it a reply to other messages, etc[6]. Each of the messages not previously classified as spam (see §3.2.) was parsed once, extracting both Header and Body data, which was stored in a database (see §3.2.2.).

The data in the body is the actual text that forms a Usenet post, and from this we counted the number of tokens for later use, but most importantly, we used a regular expression to match and count the emoticons in the text. Creating a regular expression which yielded no false positives but which did not miss any emoticons in the text at the same time proved difficult, so we opted for an approach were we used a regular expression that we allowed to match all emoticons, but which should not miss any of the emoticons in a text. This approach yielded a number of false positives addition to the list of possible emoticons (candidates), which was subsequently filtered by comparing each emoticon candidate to our previously constructed list of emoticons. If it did not occur in that list it was considered a false positive, and if it did occur we assumed it to be an emoticon and its occurrence was registered in the database, linked to the message being analyzed. We believe that this allowed us to find almost all if not all emoticons in the text, while the check for previously defined emoticons filtered out false positives. Given the large number of posts (396187), manually checking for false positives was not possible, and we are aware that a large number of false positives was stored in the database this way (see §3.2.3. for details on our accuracy). For example, the emoticons >:-) and :-) are rather similar. However, when a poster replies to a message, the message lines beloning to the messages being replied to are preceded with a > character. This, combined with a poster placing an emoticon like :-) alone on a line, will lead to that line being stored in a reply to that message as >:-). So while the first match was :-), the second time this emoticon is seen as >:-). Our classification of emoticons in the categories Positive, Neutral or Negative smoothed out many if not most of such situations, e.g. >:-) and :-) are both marked Positive, thus often resulting in the same outcome.

3.2.2. Database

Data was parsed into a MySQL database, rather than being analyzed in any on-line processing strategy. Our aim was to gather as much information from the postings while parsing them only once, and then be able to gather statistical summaries from the database without going through the files again. This is useful for repeated exploratory analysis.

3.2.3. Emoticons

We compiled a list of 2,161 unique emoticons, by combining emoticons taken from two web sources.[7] To this list we added three more: "!!!", "???" and "!?!?". These last three represented "grouping" emoticons, aggregating if strings would be encountered in a message consisting of three or more consecutive characters being either all exclamation marks, or all question marks, or a mixture of those two characters, respectively.

To allow for classification of emotion in the Usenet messages later on, the authors classified all emoticons in our collection by assigning exactly one of three possible labels to them via a web interface (see Fig. 1). The procedure for tagging the emoticons was as follows: the interface would retrieve a hitherto unclassified emoticon from the database, the user would then tag it as probably representing a positive, neutral or negative emotion by clicking on POS, ??? or NEG, respectively. The tag "???", was used both for emoticons deemed neutral as well as ambiguous by the authors. Submitting this choice would store that label for the emoticon at hand and then another, still unclassified emoticon was be retrieved for tagging. This cycle continued until there were no unclassified emoticons left. The total set of emoticons was thus tagged by two authors, but each emoticon was tagged only once. [8]

Classification of the emoticons into three aggregate classes is not a straightforward task since there is a strong argument that if anyone has bothered to type out an emoticon at all, then it is an indication of positive sentiment in the first place, even if it is not readily perceptible as a smiling face.[9] Separately, there is the issue that some of the non-textual components of the files constructed out of ASCII characters are better described as ASCII-art or flourishes than as emoticons. Examples are provided below.

We did not expect to encounter all emoticons in the database in our corpus, as some emoticons (e.g. Figure 2) seem to be meant as ASCII art rather than carriers of emotion only. Since these emoticons could be used as state-

[5] `http://spamassassin.apache.org/` — Last verified March, 2008

[6] For a detailed, technical description of the structure of Usenet messages, see the official specification (RFC1036): `http://www.w3.org/Protocols/rfc1036/rfc1036.html` — Last verified March, 2008

[7] One source was `http://www.gte.us.es/~chavez/Ascii/smileys.txt` — last verified on March, 2008. The other was `http://www.windweaver.com/emoticon.htm` — last verified on March, 2008. Both sites supplied descriptions as well as the emoticons themselves.

[8] No comparisons were carried out to check for classification consistency; however, in post classification debriefing the authors identified that the same rating strategy was used independently.

[9] We have classified frowning smiles as negative (see §3.2.3.), but a frown placed in a location that suggests sympathy is indicative of positive sentiment.

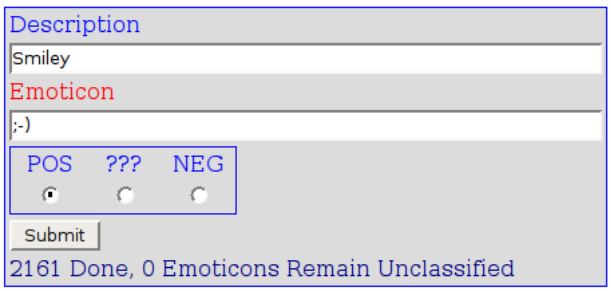

Figure 1: Emoticon Classification Interface

ments of emotion nevertheless, we classified all emoticons that we had. In the case of emoticons that were tantamount to art or graffiti, we classified them as neutral, even though their meaning could be seen as ambiguous. In cases where the emoticon's sentiment was obvious only from the accompanying text, it too was rated as neutral. In the end,the 2,164 emoticons that we recognized were divided into three groups of 668 positive, 419 negative and 1,077 neutral emoticons. Parsing the postings showed, however, that only on the order of 100 emoticons were actually used as such in the corpus.

(____(__________()~~~~~~~~~ ,=;=;=;=;=;=;=O~

Figure 2: ASCII Art: Taking a Smoke Break & Centipede.

A problem related to instances of ASCII art is the realized potential in our data set for false positives. Because we began with such a large list of emoticons, we were bound to have a large number of emoticons that were drawn from symbols actually just used as text delimiters, either automatically imposed by a news reader or inclusion of a thread of message with quoting text indicators, or through the formatting of text by a user which left in a state that had marking consistent with there being an emoticon, but for which the marking was not intended as an emoticon. Some examples of these, and demonstration from the original text that these would be realized as false positives by an automatic token seeker is provided below. In (1) and (2) one can see examples of text sequences that are plausibly construed as emoticons using > as a particular kind of hairdo, frown or smile. However, (3) and (4), respectively, show that the actual occurrences of these were unlikely to have been the intended emoticon use, where a complicating factor for (1) is that although it *is* an example of an emoticon as it occurs in our database, it is shown in (3) as the result of the concatenation of the > character denoting that this line is (part of) a text being replied to and the emoticon ":-)". So although (1) is a valid emoticon match for (3), the actual emoticon used was :-). The authors regard cases like these, too, to be false positives.

(1) >:-)

(2) =>

(3) > Charlie R wrote:
. . .

Politics			
Language	Avg. TP	Avg. FP	Accuracy
Swedish	1.6250	0.7500	0.9963
German	1.0909	0.1818	0.9996
Italian	1.7132	0.2353	0.9987
English	1.2063	0.8730	0.9974
Science			
Language	Avg. TP	Avg. FP	Accuracy
Swedish	0.3636	2.0000	0.9835
German	1.0870	0.8478	0.9943
Italian	1.3478	0.4348	0.9951
English	1.3529	0.4118	0.9951

Table 1: Average True and False Positives & Accuracy for 400 Random Samples of Postings

>:-) men du får hålla med om att det är någonting sjukt med dessa

(4) Re: => Bu$h LIED about Rumsfeld - GAVE ELECTIONS to the DEMOCRATS <= thank's idiot chimp!

Given the scale of the data set even with just two topic areas and four languages as reported in Table 2, it is impossible to set about removing false positives classifications from the counts reported below. However, we can provide an estimate of the False Positive Rate as a measure of error associated with the frequencies and averages that we do report. To do this we randomly sampled 400 messages (approx. 0.1%) from the messages in which our parser (described in §3.2.1.) found at least one emoticon. The samples were taken of each of the eight categories and report the average True Positives (TP), the average False Positives (FP) and the average of a composite accuracy rate as described in (Fawcett, 2006). We wanted to use stratified random sampling, but given the large differences in numbers that make up the largest and the smallest categories, we feared that the smallest categories would be under represented, or perhaps not represented at all in a random selection of postings. To prevent underrepresented categories, our sampling strategy was to start by randomly selecting 20 samples from each language, evenly divided over "Politics" and "Science". In addition to these 80 samples, 320 more where chosen at random without language and topic constraints.
The results, ordered by language and topic, are in Table 1, where the cell-contents indicate the average true and false positives (TP & FP) and the accuracy(Fawcett, 2006).

3.3. Description of the Data

To give an idea of the size and contents of our corpus, we present some summarizing statistics. To make comparisons possible across languages and topic areas, it is necessary to keep track of the total traffic volume; this is depicted in Table 2. This is included here in order to provide a picture of the scale of the data at stake (after spam has been removed), and to demonstrate the uneven balance of postings across categories. More fine grained descriptive statistics are provided in §4.1.

It is also useful to have a general image of interactivity across the categories analyzed, before identifying how use of emoticons (positive, negative or ambiguous) varies across those categories. One way to measure interactivity is in terms of the number of distinct posters responsible for the postings in the category. Thus, we record, in addition to the number of postings, the average number of postings per individual (APPI). See Table 2.

Language	Topic	Messages	APPI	ANR
Swedish	Politics	18225	23.13	0.2177
Swedish	Science	814	5.73	0.1818
Sum:		19039		
German	Politics	933	3.30	0.1565
German	Science	75230	12.72	0.0988
Sum:		76163		
Italian	Politics	173672	32.94	0.0986
Italian	Science	32117	5.97	0.0908
Sum:		205789		
English	Politics	81635	10.90	0.2107
English	Science	13561	10.66	0.1036
Sum:		95196		
Overall Sum:		396187		

Table 2: Messages per language per topic

Swedish and Italian political discussions appear to involve the greatest levels of interactivity on this measure. German political discourse shows the least interactivity and Italian and Swedish science discussion show quite little. These comparisons are relative to the English baseline, where interactivity between science and politics turns out equal. However, this is a coarse grained metric of interactivity as it does not account for whether posters have read each other's texts, only the volume of traffic and posters.

A separate measure of interactivity is in cross postings — a cross-posting may become visible from one group to another for a reader who reads the second only, and not the first. Thus, cross-postings have an effect of increasing visibility of postings. This is not a value that we recorded in this experiment. We are not at present certain how to take mere visibility into account in measuring interactivity. There is a binary distinction for each posting: it is either a new posting or a reply. This makes it useful to track the average number of new postings per poster (which is a slightly less interactive occupation than replying to an existing posting) and also the average number of reply postings per poster. This number provides a useful view on interactivity within a group. The data here, shown in Table 3, indicates that there are in general substantially more postings that are new than there are replies: when the ratio exceeds the value one, then there is relatively low direct interactivity in terms of posters replying to existing postings. However, this does not mean that the postings are in complete isolation. Individuals can be reading messages and sending replies to them without marking them through the system as replies. Nonetheless, those sorts of replies are not directly measurable, as again, that sort of interactivity hinges on greater depth of textual analysis to uncover artifacts of visibility of instigating messages within follow-on postings that are not replies.

Language	Subject	NewPosts	Replies	NP/R Ratio
Swedish	Politics	11599	6626	1.75
Swedish	Science	769	45	17.09
German	Politics	580	353	1.64
German	Science	45080	30150	1.50
Italian	Politics	133592	40080	3.33
Italian	Science	20498	11619	1.76
English	Politics	60276	21359	2.82
English	Science	10515	3046	3.45

Table 3: Ratio of New Postings to Replies

Here, again, Swedish and Italian have markedly different values for this measure between discussions on politics and science — within Swedish postings, they differ by about a factor 10, while in Italian the difference is only nearly a factor 2. No obvious pattern emerges. If there are patterns, then they are not detectible with this metric in our dataset (which could be too small for such discoveries in terms of comparisons of both languages and topics). Although the NP/R outlier of 17.09 in Swedish Science postings is remarkable, this could possibly be due to the small number of posts in that category, combined with a small number of enthusiastic thread-starters.

A final measure of interactivity that we address is examining the average number of messages referred to (ANR) in any message, as shown in the ANR column in Table 2. This is operationalized from the message header rather than from the body of the message: a fully new message can easily make reference to a message that "someone posted about a year ago" without it being a reference that is tracked in the header of the posting by the news-reader.[10] The posting references that are recorded in the message header partially indicate an amount of the discussion "thread" at the point of writing. It is only partial because someone can reply to a note at any point along in the thread, creating branches, and users can also interfere with the information recorded there. Another caveat is that the official protocol[11] used by ISP's to exchange Usenet messages states that "It is permissible to not include the entire previous 'References' line if it is too long. An attempt should be made to include a reasonable number of backwards references." We see no way of knowing how many references have been removed[12] by ISP's without doing a full text analysis, which is beyond the scope of this experiment.

Thus, the information isn't a complete picture of thread interactivity, however, it does give some indication of how

[10] Conversely, a message that is a reply to a posting may be marked as such in the header, but not include mention of the prior posting anywhere in the message body.

[11] For a detailed, technical description of the structure of Usenet messages, see the official specification (RFC1036): `http://www.w3.org/Protocols/rfc1036/rfc1036.html` — Last verified March, 2008.

[12] The largest number of references observed in our data was 30.

long conversations on a topic last before new threads are created. The fact of a thread existing is less rich a notion of interactivity than an estimate of average thread length for a group. Longer threads are more interactive because they often reflect group discussion rather than single individual comment and reply sequences (although threads can consist of strictly dialog as well as n-alog).

Table 4 shows the average number of word-level tokens per posting per language. Again, we do not think it possible to reliably determine the interactivity between posters by comparing the average tokens per message to the amount of added tokens per reply without doing a full text analysis, because users are free to edit and remove (parts of) the texts that they reply to. Still, Tables 4, 5 and 6 are provided to show some differences between posts given their language and topic in our corpus.

We attempt to provide a broad view on interactivity, so we comment on the relation between interactivity and emoticon use. This is done by reporting on two types of correlations, each based on an interactivity measure discussed previously. Section 4.4.2. discusses the correlations between the number of messages referred to in messages (see §3.3. & Table 2) and the three types of emoticon categories (positive, negative, & ambiguous), followed by a similar discussion but now based on the correlations between the message length, measured in tokens[13], and the three emoticon types.

Language	Avg. Tokens per Message
Swedish	253.84
German	193.70
Italian	166.92
English	275.03

Table 4: Avg. Tokens per Message per Language

Topic	Messages	Avg. Tokens per Message
Politics	274465	208.51
Science	121722	188.04
Overall Sum:	396187	*Overall Avg.*: 202.2218

Table 5: Messages & Avg. Tokens per Message per Topic

Language	Topic	Avg. Tokens per Message
Swedish	Politics	249.34
Swedish	Science	354.44
German	Politics	270.99
German	Science	192.74
Italian	Politics	163.80
Italian	Science	183.81
English	Politics	293.80
English	Science	162.03

Table 6: Avg. Tokens per Message per Language per Topic

Language	Emoticons	No Emoticons	% With
Swedish	4064	14975	21.3%
German	21294	54869	28.0%
Italian	46931	158858	22.8%
English	18327	75869	19.5%

Table 7: Number of Postings With and Without Emoticons

4. Results & Discussion

4.1. General findings

Swedish, Italian and English do not make significantly different use of emoticons in terms of the percentage of postings with them, as Table 7 shows. German had relatively more postings with emoticons than the other languages.

Among the postings that had emoticons in them, Table 8 shows the average number of emoticons per posting, as well as the standard deviation. A factor that would interact with the number of possible emoticons per posting is the average message length.[14] The standard deviations reported in Table 9 demonstrate that the data is skewed on this measure. Several of the emoticons we have observed may have been counted in the other posts of that thread. As messages can include entire posting histories, longer messages with many messages included from a thread of communication are more likely to contain emoticons than shorter ones. Table 10 shows the averages and standard deviations in terms of tokens per message with all postings, so including messages without Emoticons.

As message length is comprised of two variables, one being how much an individual adds to the length of a message, and the other how many (parts of) messages on average are included in replies, it is hard to draw conclusions in terms of interaction from these figures.

4.2. Differences Per Language

Table 11 shows that just under half the emoticons used by the Swedish writers were externally classified as positive, while exactly half of the Italian emoticons were negative. The remainder of the Swedish texts were evenly split between negative and ambiguous categorizations, while two thirds of the remainder for the Italian writers were positive, and only 16% ambiguous. The German data patterned roughly with the Swedish data: a preponderance of positive emoticons (65% of those used), with the remainder essentially evenly split between negative an neutral. The English texts showed the same trend but with only 40% of the emoticons being positive.

The results in Table 12 show that the number of positive, negative and neutral emoticons relativized to total number of postings was not significantly different for Swedish and English. Interesting divergences appeared for German, where there are twice the ratio of positive emoticons to

[13] We use the term "token" for sequences of letters of the alphabet individuated by spaces, line termini, or punctuation. Emoticons were not considered to be tokens.

[14] We found no convincing correlations between use of emoticons (positive, negative or neutral) and message length, cf. §4.4.2.

Language	Average	St. Dev. (σ)
Swedish	1.5182	2.9755
German	1.5061	1.5899
Italian	1.5563	1.5563
English	1.6548	1.6548

Table 8: Frequency of Emoticons per Posting with at Least One Emoticon

Language	Average	St. Dev. (σ)
Swedish	354.2	630.4
German	279.2	1332.6
Italian	338.7	751.2
English	538.7	1513.9

Table 9: Average Tokens per Message with at Least One Emoticon

Language	Average	St. Dev. (σ)
Swedish	226.6	408.0
German	184.2	724.5
Italian	262.3	436.9
English	307.5	794.5

Table 10: Average Tokens per Message Overall

Language	Positive	Negative	Ambiguous
Swedish	0.46	0.27	0.27
German	0.65	0.16	0.19
Italian	0.34	0.50	0.16
English	0.40	0.30	0.30

Table 11: Ratio of Emoticon Type to Total Emoticons, by Language

postings as in Swedish and Italian, while the ratio of negative emoticons per posting is twice that of Swedish.

4.3. Differences Per Topic

Analyzing the differences by topic reveals that many of the within language differences may have explanations that hinge less on inter-cultural differences in comment and feedback mechanisms. Table 13 and Table 14 illustrate this. Over half of the emoticons used in science discussions in Italian are positive, but a greater majority of the emoticons used in politics discussions are negative. In the case of science, the remainder is evenly split, and in the case of Italian, two-thirds is positive and the final portion is ambiguous. In German, politics is more or less evenly divided across the categories, but in science, the emoticons are 65% positive. The English pattern is much the same as the German pattern. The Swedish pattern for science is overwhelmingly (67%) towards ambiguous emoticons, with a roughly even split between positive and negative for the remainder; however, for politics the Swedish emoticons are just under half (48%) positive, with the remainder evenly split. These measures are all based on the distribution of positive negative and ambiguous emoticons among the emoticons actually used. It remains to consider those distributions relative to the number of postings in each category: Tables 15 and 16 are the counterparts of Tables 13 and 14 but are derived from the data summarized in Table 12 rather then Table 11. If the tables are compared per topic, we can see that they show the same patterns, and the differences seem to be in scale only, although the scales do differ a bit between groups. The rows displaying the data for English show the only remarkable difference between the topics.

4.4. Differences Per Levels of Interactivity

4.4.1. Number of Posters

Table 2 indicated by language and topic area what the average number of postings per news poster is. This supplied one metric of activity within a newsgroup. The higher the average, the more involved the posters are. The Italians (32.94 messages per poster) and Swedes (23.13 messages per poster (mpp)) were most active in discussing politics. The English had the same average number of postings in politics and in science (just under 11 mpp). Separately, we have shown the use of positive, negative, and ambiguous emoticons by language and subject area. The Germans were least active in posting on politics (3.30 mpp). Recall that the German and English use of emoticons in politics was balanced across the three categories. Swedish and Italian both demonstrated high activity in politics, and the Swedish postings tended towards positive emoticons (48% of them), while the Italian postings tended towards negative (57%). This seems to represent a bona fide inter-cultural difference in emoticon use in political discussion.

In scientific newsgroups there was not a significant difference in average postings per person betwen the English (10.66) and the Germans (12.72), although the German rate indicated the highest level of activity. Recall that the German also exhibited the largest proportion of positive emoticon use (65%) in this category, and English followed closely behind (60%). The main difference between German and English was that the remainder for English focused on negative emoticons at twice the rate of ambiguous ones (26% to 14%) while the german data was more evenly balanced (16% to 19%). Italians (5.97 mpp) and Swedes (5.73 mpp) had the least activity in this subject area, but a comparable amount. Again there was a stark difference. Here, the Italian emoticons were more than half (53%) positive and the remainder closely split between negative (25%) and ambiguous (22%), while Swedish emoticons were mostly ambiguous (67%), with the remainder closely divided between positive (14%) and negative (19%). The generalization appears to be that the German and English postings are about the same in their use of emoticons, while Swedish and Italian postings differ sharply. Italian postings exhibit distinctively positive emoticon distributions for scientific discussion and negative emoticon distributions for political discussions. Swedish postings tend towards the positive in political discussions, but tend to be overwhelmingly more ambiguous in scientific discussions. All of these considerations, of course, are relativized to the particular

Language	Positive	Negative	Ambiguous
Swedish	0.18	0.11	0.10
German	0.36	0.09	0.10
Italian	0.15	0.22	0.07
English	0.16	0.12	0.12

Table 12: Ratio of Emoticon Type to Total Postings, Relativized to Total Postings, by Language

Science			
Language	Positive	Negative	Ambiguous
Swedish	0.14	0.19	0.67
German	0.65	0.16	0.19
Italian	0.53	0.25	0.22
English	0.60	0.26	0.14

Table 13: Ratio of Emoticon Type to Total Emoticons, by Language, for Science

corpus that we tested on. One can imagine that politics is sensitive to the temporal span of testing, in any case. The question is how the findings here would transfer to other temporal spans, or if this one contained unique features that would spark positive feelings in Sweden and negative ones in Italy, and such that the results are not transferable.

4.4.2. Exchange Length

Our final measurements were aimed at identifying correlations between interactivity and the use of emoticons. Here we consider two different ways of measuring levels of interactivity and consider the correlation between overall emoticon use and the interactivity measures. We show first the relation between each, and overall emoticon use and the use of emoticons within each of the three levels of polarity. We then show the correlation between the two measures of interactivity in terms of length of exchange, measures which prove to be independent.

Because emoticons were not counted as tokens (see §3.2.1. and Tables 8 & 9), the message length in tokens and the number of references in messages were both treated as independent variables with respect to the emoticon counts.

The correlation between message length as measured by token count and total use of emoticons was 0.090, ($p < 0.0001$, two-tailed). The correlation between the number of positive emoticons ($N = 50017$) in a message and the message length was 0.093 ($p < 0.0001$, two-tailed). For the negative emoticons ($N = 39753$) there also is a very weak correlation, $r = 0.078$, between message length and emoticon count, but still significantly so ($p < 0.0001$, two-tailed). The strongest correlation ($r = 0.151$), a very weak positive correlation, can be reported for neutral emoticons ($N = 11808$) when paired with message length ($p < 0.0001$, two-tailed). Thus, message length has no overall correlation with emoticon use, but a weakly positive correlation with use of neutral emoticons.

Using the other index of interactivity, the number of references to other messages retained in a posting's header, we found an insignificant ($p < 0.1346$, two-sided) marginally

Politics			
Language	Positive	Negative	Ambiguous
Swedish	0.48	0.28	0.24
German	0.34	0.34	0.31
Italian	0.28	0.57	0.15
English	0.37	0.31	0.32

Table 14: Ratio of Emoticon Type to Total Emoticons, by Language, for Politics

Science			
Language	Positive	Negative	Ambiguous
Swedish	0.08	0.12	0.40
German	0.36	0.09	0.10
Italian	0.33	0.16	0.14
English	0.17	0.07	0.04

Table 15: Ratio of Emoticon Type to Total Postings, by Language, for Science

negative correlation with overall use of emoticons ($r = -0.004$). Within that, for the number of positive emoticons and the number of messages referred to we found $r = 0.075$, a marginally positive correlation. Use of negative emoticons had a negligible negative correlation ($r = -0.009$) with number of references that did not reach significance ($p < 0.031$, two-tailed). Finally, neutral emoticons had a similar correlation with the number of references, but with even less significance ($r = -0.004$, $p < 0.3468$, two-tailed). On this measure of interactivity, as well, there is essentially zero correlation between emoticon use and interactivity.

The two measures of interactivity are overall message length as determined by the number of alphabetic tokens (that is, exclusive of emoticons), and the number of prior messages referred to in the posting headers. The correlation between these two variables was 0.008 ($p < 0.0070$ two-tailed). The significant lack of correlation suggests that these two variables are independent, and thus cannot stand proxy for each other as indices of interactivity. We conclude that there is no correlation of interest in the comparisions we made, and therefore that emoticon use is not a suitable indicator for interactivity given either of these independent metrics. It remains to examine the actual patterns of interaction more explicitly to determine if a substantial relationship exists there. However, if the results for these two independent metrics of interactivity transfer, then one might be inclined to conclude that emoticon use does not increase as a function of interactivity in communication, except for the use of neutral emoticons where there is a weak positive correlation with length of contribution in this style of electronic communication.

5. Conclusions & Further Work

This paper represents a small contribution to research involving emoticons. Much other work uses emoticons to classify overall sentiment of documents (Read, 2005;

Politics			
Language	Positive	Negative	Ambiguous
Swedish	0.18	0.11	0.09
German	0.12	0.12	0.11
Italian	0.11	0.23	0.06
English	0.15	0.13	0.13

Table 16: Ratio of Emoticon Type to Total Postings, by Language, for Politics

Suzuki et al., 2006; Neviarouskaya et al., 2007) or to associate emoticons with words in texts (Nicolov et al., 2008). The work reported here are our initial experiments in attempting to analyze inter-cultural differences in emoticon use. Clearly, this is just the tip of the iceberg in analyzing inter-cultural differences in expressing sentiment using emoticons in on-line asynchronous communication.

Other forms of sentiment would also be very interesting to monitor, such as lexicalized and idiomatic politeness markers, cross-linguistically. We have seen that a major pitfall that such research faces is in properly aligning newsgroups across languages into the same topical domain.

To increase the accuracy when reporting results of research like this, more work on the analysis of interactivity needs to be carried out. We are thinking along the lines of token count variance in threads, as that could indicate how much new tokens are added and cited with each reply, but we are also considering a method tailored to individual posters where we analyze how often a poster starts a new thread (not interaction *yet*, but a necessary first step), how often on average a poster replies to treads in which he has not yet participated (moderate interaction). Also, it could be interesting to include time stamp analyses of posts to get a measure of interactivity; less time between posts could indicate a more lively debate. A final measure could perhaps make use of data on how often average posters actively engage in discourse by replying more than once to the same thread, possibly taking into account who take part in the discussions; this is based on the thought that multiple interactions with the same people can be more social than multiple brushes with new individuals all the time.

The authors would like to add to the methods used to analyze postings more accurate parsing methods, allowing at least for the separate parsing of the parts of the message(s) that form cited parts of postings versus the added response(s) by a poster. It will probably prove very difficult to keep track of the origins of the lines forming a cition from previous messages, since even with access to a complete collection of postings, there is always the added difficulty that individual users could have edited "cited" lines, making the tracing and recognition proces inaccurate and time consuming.

A different avenue potentially worth pursuing, related to interactivity but explicitly acknowledging external influences, could take the shape of a cross cultural comparison of responses to, for instance, current events made available by (online) news sources.

6. Acknowledgements

This research is supported by Science Foundation Ireland RFP 05/RF/CMS002.

7. References

Irene Albrecht, Jörg Haber, and Hans-Peter Seidel. 2002. Automatic generation of non-verbal facial expressions from speech. In *Advances in Modelling, Animation and Rendering*, pages 283–293.

Jens Allwood and Loredana Cerrato. 2003. A study of gestural feedback. In Patrizia Paggio, Kristiina Jokinen, and Arne Jönsson, editors, *Proceedings of the First Nordic Symposium on Multimodal Communication*, pages 7–22.

Jens Allwood. 1985. Intercultural communication. In Jens Allwood, editor, *Tvärkulturell Kommunikation, Papers in Anthropological Linguistics 12*. University of Göteborg. English translation by Jens Allwood.

Jens Allwood. 1999. Are there swedish patterns of communication? In H. Tamura, editor, *Cultural Acceptance of CSCW in Japan & Norid Countries*, pages 90–120. Kyoto Institute of Technology.

Loredana Cerrato. 2003. A comparative study of verbal feedback in italian and swedish map-task dialogues. In *Proceedings of the Nordic Symposium on the Comparison of Spoken Languages,*, pages 99–126.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2007. Narrowing the social gap among people involved in global dialog: automatic emotion detection in blog posts. In *Proceedings International Conference on Weblogs and Social Media*, pages 293–294. Omnipress: Boulder, Colorado, USA. ICWSM-07: Poster Paper.

Nicolas Nicolov, Franco Salvetti, and Steliana Ivanova. 2008. Sentiment analysis: Does coreference matter? In *Symposium on Affective Language in Human and Machine*. AISB: Aberdeen, UK.

Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics.

Yasuhiro Suzuki, Hiroya Takamura, and Manabu Okumura. 2006. Application of semi-supervised learning to evaluative expression classification. In *Computational Linguistics and Intelligent Text Processing*, pages 502–513. Springer: Berlin / Heidelberg.

Universal or Culture-Specific Metaphors in Economics? A Corpus Study of Original vs Translated Italian

Maria Teresa Musacchio
Dipartimento di Lingue e Letterature AngloGermaniche e Slave, Università di Padova
Via Beldomandi 1, 35137 Padova (Italy)
mt.musacchio@unipd.it

Abstract

This paper compares the use of metaphor in a corpus of Italian economic reports (standard sources of economic information), and Italian newspaper and magazine articles and articles from *The Economist* translated into Italian (popular economics). The aim is to identify differences in the frequency and types of metaphors in reports as opposed to articles and in original Italian versus translated articles with a view to detecting text-typological and/or culture-specific features of metaphor-building in Italian economics. First, wordlists of the three sub-corpora are created; second, keywords are extracted using economic reports as the reference sub-corpus and KWIC concordances are run to investigate differences in the use or kinds of metaphors in the two text types analysed. Third, further candidate metaphors were studied on the basis of wordlists reflecting typical world hypotheses in economics. Finally, concordances of keywords and further candidate metaphors were run to investigate if there are culture-specific metaphors in the three sub-corpora.

1. Introduction

Over the centuries, several approaches to metaphor have been proposed. In classical times, Aristotle described metaphor as a kind of substitution, that is a transfer of meaning from one expression to another. The idea also developed that metaphor was essentially an implicit comparison between two elements. In modern times, Richards (1965) gave new impetus to metaphor studies by distinguishing three aspects of metaphor, the *vehicle* or metaphorically used item, the *tenor*, i.e. the metaphorical meaning of the vehicle, and the *ground*, that is the common elements of meaning used as a basis for the metaphor. This gave rise to the interactive model of metaphor, which implied that metaphor can convey knowledge but that cognitive meaning cannot be retrieved through a literal 'translation' of the metaphor. The idea that there must be some connecting elements between the tenor and the vehicle had also been developed by Black (1962), who however rejected both the substitution and the comparison views and described metaphor as a projection of associative implications from one entity to another. Within relevance theory metaphor has been regarded as interpretive use of language where meaning can be inferred on the basis of relevant resemblances in context (Sperber and Wilson 1986/1995). Lakoff and Johnson (1980/2003) consider metaphors as essential conceptual components of human cognition which involve a source domain, a target domain and a set of correspondences. Correspondence can be 'ontological', i.e. involving elements in the source and target domains, or 'epistemic', that is involving relations about the elements. Lakoff and Johnson further identified 'elaborations' or more specific versions of a basic metaphor, and 'metaphorical entailments' or patterns of reasoning that carry over from the source to the target domain (Cruse 2004: 198-207). In terms of functions of metaphors, Kövecses (2002) further distinguished structural and orientational metaphors – which differ from ontological metaphors as they respectively provide a richer and poorer conceptual structure for target domains.

The debate on the nature and understanding of metaphorical meaning has also carried over in science where exegetical or pedagogical metaphors are regarded as playing a role in teaching or explaining theories, while constitutive metaphors are – at least for some time – considered as an irreplaceable part of a scientific theory (Boyd 1979: 359-60). With reference to science, Kuhn (1979: 414-5) drew a useful distinction between a metaphor and a metaphor-like process and further pointed out that long after the process of exploring the potential similarities or analogies between the source and target domains has ended, a metaphor can remain essential to a theory to the point that the metaphor-like process is never completed.

In economics, Henderson's (1982, 1993) and McCloskey's (1988, 1995) view of metaphor and the metaphorical process is elaborated in Klamer and Leonard (1993) who identify three types of economic metaphors, pedagogical, heuristic and constitutive metaphors. Pedagogical metaphors coincide with exegetical or pedagogical metaphors in science, while heuristic metaphors serve to catalyze thinking or propel thought. They originate from constitutive metaphors that provide the necessary conceptual schemes through which economists interpret a world "that is either unknowable (…) or at least unknown (Klamer and Leonard 1993: 39). In other words, constitutive metaphors determine what makes sense and what does not in economic terms. On the basis of work by Foucault and Pepper, Klamer and Leonard further identify four world hypotheses – organicism, mechanism, formism, and contextualism. These hypotheses are relevant as they are characterized by different root metaphors that can be seen to play an important role in economics. For example, organicism means that the economy is seen as a living thing, mechanism is found to underlie the ideas of the price mechanism, equilibrium and elasticities, formism stresses the organization of the world, while according to contextualism economics has a history in which events are contiguous and human actions can only be understood in context.

When discussing any feature of science, including metaphor in economics, consideration of text type is essential – at least the usual tripartite classification in academic, instructive and popular-science texts. Indeed, Hundt (1998: 109) points out that popular economics texts

differ from other text types as they contain explanations of concepts, their metaphors are much more extended than in specialized texts, and they exhibit a simpler syntax than academic or institutional texts. When referring these remarks specifically to metaphor-building, one could infer that in popular economics texts as – more generally – in popular science, some conceptual metaphors are culture-specific since they arise "both from cognitive, bodily constraints and from shared experience" (Eubanks 1999: 181).

This paper compares the use of metaphor in a corpus of Italian economic reports (standard sources of economic information), and Italian newspaper and magazine articles and articles from *The Economist* translated into Italian (popular economics). The aim is to identify differences in the frequency and types of metaphors in reports as opposed to articles and in original Italian versus translated articles with a view to detecting text-typological and/or culture-specific features of metaphor-building in Italian economics.

2. Aims and Methodology

As has been outlined above, relevant contributions to the contemporary study of metaphor have come from cognitive linguistics, psychology, rhetoric and philosophy. Scientists have also helped to throw light on the functions of metaphors and metaphor-like processes either in science in general (e.g. Kuhn) or in their own discipline (e.g. McCloskey and Henderson for economics). Cognitive approaches such as Lakoff and Johnson's assume that metaphors are conceptual and hence originate from central processes and structures of human thought. This implies that they are not language-specific but rather universal. Cross-linguistic research to investigate the possibility that metaphors are not language-specific has shown that at least some of them are shared. However, cases were detected were no complete consistency could be found. In particular, studies of metaphors in economics texts in English, French and Dutch have shown differences that were ascribed to cultural factors (Deignan and Potter 2004: 1232-3).

Drawing on these various strands of research, the present study aims to compare metaphors in two text types – economic reports and magazine or newspaper articles – and contrast the use of metaphorical processes in articles originally written in Italian and articles translated from English into Italian. The aim of the former investigation is to detect any differences in the use of metaphors as a function of the purposes which different text types may serve. (This inquiry is essentially monolingual and covers two text types that – unlike academic papers – are still quite commonly written in Italian, i.e. reports as sources of economic information and magazine/newspaper articles as popular economics texts.) The latter investigation has the objective to identify alternative metaphors and/or different frequencies in use in original as opposed to translated Italian articles. This study is to a certain extent cross-linguistic since it takes into account that differences in use may derive from an influence of the source language – in this case English – on the target texts.

First, in order to identify metaphors that could be worth studying, wordlists of the three sub-corpora were used to obtain two sets of keywords, one for original Italian articles and the other for translated articles. In both cases keywords were extracted using economic reports as the reference sub-corpus. This part of the investigation has been based on the assumption that economic reports as the more specialized texts use the standard special language of economics including metaphors. Second, once keywords had been extracted that could point to candidate metaphors in original and translated Italian articles, KWIC concordances were run in the three corpus components to investigate use in greater detail. Third, further candidate metaphors were searched for in the corpus by building lists of words and/or terms that are related to the world hypotheses economics generally draws on in metaphor-building – physics, health and medicine, change, war and weather – also following Resche's (2000, 2004) approach in her study of the terminology of economics. Finally, for the key words and terms, KWIC concordances were run to test whether they did form culture-specific metaphors in the three corpus components. Corpus analysis was conducted using version 5 of the WordSmith Tools software developed by Mike Scott.

3. Data Collection

The corpus is divided into three sub-corpora: a) a collection of economic reports published by the Bank of Italy and ISTAT, the Italian National Institute of Statistics, plus speeches delivered by the former Governor of the Bank of Italy (which can be regarded as written texts as they usually read, not delivered); b) articles from the Italian economic and financial daily *Sole 24 Ore*, from the economic pages of the national Italian daily *Corriere della Sera* and from the Italian business weekly *Economy*; c) articles from *The Economist* and *The World In* series published by The Economist Publications and translated into Italian and published by *Economy* and in special supplements of the Italian daily *La Stampa* respectively.

Text type	**Source**	**Tokens**	Docs
Reports + speeches	Bank of Italy, ISTAT	185,437	7
Original Ital. articles	*Economy*, *CorSera*, *Sole24Ore*	63,920	71
Translated Ital. articles	*Economy*, *LaStampa* Suppl.	98,207	105
TOTAL		**347,564**	**183**

Table 1: Components of the corpus of economics grouped into the three sub-corpora of economic reports, original Italian articles and translated Italian articles.

The corpus currently has 347,564 tokens distributed amongst 183 texts. The average number of words per document varies from 26,491 (economic reports) to 900.3 (original Italian articles) and to 935.3 words per document (translated Italian articles). This small corpus covers the period from 1999 (economic reports) to 2008 (part of the original Italian articles) and is still being slowly developed, since translated articles are scanned because they are not available in electronic format while original

ones are chosen to match the topics covered in the translated articles and keep the corpus balanced. However, at the moment the corpus does include virtually all the *Economist* articles published in translation by *Economy* as long as it had exclusive copyright for Italy.

4. Analysis

Keyword lists highlighted different 'keyness' in the two text types of terms describing change: *dinamica, andamento, crescita* and *flessione* were keywords in economic reports, *rialzo* and *ribasso* in newspaper or magazine articles. These terms describe a change in economic conditions which is a central concept in economics. The pattern of variation is described metaphorically in Italian by using the sub-technical terms *dinamica/dinamiche* (deriving from the physical term dynamics) and *andamento/i* (trend). The two key metaphors for increase (Lakoff and Johnson's UP metaphors) are *crescita* (growth) – based on the idea of the economy as a living thing – and *rialzo* (lit. a rise, but also a wedge). Metaphors for decrease (Lakoff and Johnson's DOWN metaphors) are *ribasso/i* (reduction, rebate) and *flessione/i* (lit. flexion, bending – a physical term).

In Table 2a bold-type marks the highest-frequency metaphorical terms. As can be seen, the standard economic terms (*dinamica, andamento, crescita, flessione*) are drawn from the emotionally neutral language of science and especially of physics and are most frequent in economic reports, while the 'journalistic' terms *rialzo* and *ribasso* are more common in original Italian articles. *Dinamica* and *andamento* are least frequent in translated articles – possibly because they are not one-to-one equivalents of *pattern* or *trend.*

Word/term	Reports+speeches	Orig.Ital.	Transl.Ital.
dinamica/he	**1.2**	0.3	0.02
andamento/i	**0.6**	0.3	0.03
crescita	**3.4**	2.1	2.1
rialzo/i	0.3	**0.8**	0.2
ribasso/i	0.1	**0.2**	0.1
flessione/i	**0.8**	0.2	0.1

Table 2a: The distribution of the keywords related to change given as frequency per thousand words of text. Where applicable, relative frequency is calculated using the sum of noun occurrences in the singular and plural forms.

The KWIC concordance in Table 2b below shows how words indicating change are used in economic reports. As is often the case in economics, words are used both in their general language meaning and as terms or LSP collocations. Consider *dinamica,* which collocates with terms such as investment, production, prices and consumption, but is also used to refer to the trend of the business cycle (*dinamica congiunturale*). Further, as they only describe a pattern metaphorically, *dinamica* and *andamento* collocate with formal adjectives such as *sostenuto* (high), *moderato* (moderate), *favorevole* (positive), and *sfavorevole* (negative) that attach a degree of 'sentiment' to data and thus guide readers in the interpretation. In this case, guidance is provided with reference to evaluation or judgment.

LH co(n)text	Term	RH co(n)text
quelle in conto capitale sono cresciute per effetto della sostenuta	**dinamica**	degli investimenti (10,7 per cento) e della ripresa dei pagamenti
La creazione di posti di lavoro a fronte di una contenuta	**dinamica congiunturale**	della produzione conferma il mutamento
e dallo sfavorevole divario nella	**dinamica dei prezzi,**	l'espansione del prodotto è scesa
per il secondo anno consecutivo. La vivace	**dinamica dei consumi privati,**	che nel 1998 sono cresciuti del 4,8%,
agli eccezionali proventi connessi con il positivo	**andamento**	dei mercati finanziari, difficilmente ripetibile.
i saldi complessivi hanno beneficiato anche di un	**andamento**	favorevole dei tassi di interesse.
Il comparto della pesca ha scontato di nuovo un	**andamento**	sfavorevole in termini di quantità prodotte
all'arrotondamento delle cifre decimali. La	**flessione**	dell'avanzo primario riflette soprattutto l'andamento delle entrate,
a quasi quattro volte quello dell'intero saldo finanziario del settore. La	**crescita**	si è attenuata nella seconda metà dell'anno, in concomitanza con
pagati dagli investitori esteri. Alla	**flessione**	dei flussi di capitali privati si è contrapposto un netto aumento
la crescita del PIL nel 1998 e nel 1999 e la	**flessione**	dei tassi d'interesse, più accentuata del previsto.

Table 2b: A concordance of *dinamica, andamento* and *flessione*. [Sub-corpus: R (Reports and Speeches)].

In original Italian articles upward and downward changes are termed *rialzo* and *ribasso* respectively (cf. Table 2c). Here we can see how a successful metaphor engenders a metaphor-like process that gives rise to compounds (*revisione al rialzo/al ribasso,* literally translated: upward/downward adjustment) and derivatives (the verb *rialzare* and the adjective *rialzista*). Moreover, in newspaper and magazine articles metaphors appear to be freely mixed: *revisione al rialzo* is followed by the metaphorical *andamento* and *rialzare* – in the collocation *rialzare i tassi* (to increase rates) is used, literally, 'to throw water on the fire of prices and recovery (*per gettare acqua sul fuoco dei prezzi e della ripresa*), or to pour oil on the troubled waters of prices – and recovery.

LH co(n)text	Term	RH co(n)text
ha indotto fin da subito una significativa	**revisione al rialzo**	Delle previsioni sull'andamento del Pil nel 2007
la produzione industriale deve ancora mostrare un sostenuto	**trend rialzista",**	prosegue il rapporto. Secondo l'Outlook sebbene la fase
è probabile che la Fed si troverà ben presto nella necessità di	**rialzare**	i tassi per gettare acqua sul fuoco dei prezzi e della ripresa.
dell'1,1% per il 2003 e del 2,3% per il 2004.La	**revisione al ribasso**	è ampia per tutti i grandi paesi di Eurolandia.

Table 2c: A concordance of *rialzo, ribasso* and their derivatives or compounds [Sub-corpus: O (Italian originals)].

Change is a key concept in business cycles too. Stages in the business cycle are termed by drawing variously from physics and medicine with frequencies as shown in Table 3a: *espansione* (expansion), *contrazione* (contraction), *rallentamento* (slowdown); *crisi* (crisis), *depressione* (depression), *ripresa* (recovery). Here the same metaphors are used in Italian and English. In Italian, however, a distinction is made between the *ciclo economico*, which usually refers to change over the medium or long term, and *congiuntura* (lit. joint; circumstance) or short-term change.

Word/term	Reports+speeches	Orig.Ital.	Transl.Ital.
ciclo/i (economico/i)	0.2	**0.3**	0.1
congiuntura/e	0.2	**0.4**	*0.01*
espansione	**0.3**	0.1	0.2
boom	-	0.1	**0.4**
crisi	**1.0**	0.11	0.6
contrazione	**0.5**	0.1	0.04
recessione	0.1	**0.3**	**0.3**
depressione	-	0.03	0.02
ripresa	0.5	**1.1**	0.6
rallentamento	**0.7**	0.3	0.3

Table 3a: Metaphorical terms used to describe stages in the business cycle. Highest frequency – per 1,000 words – is marked in bold.

Both in economic reports and in the press *ciclo economico* and *congiuntura* are often used as synonyms. C*ongiuntura* is least frequent in translated articles since it has no direct equivalent in English. In the concordances from economic reports (Table 3b) the terms form noun + adjective or noun + preposition + noun combinations but two metaphors are mixed in *contrazione del ritmo di crescita* (lit. contraction of the pace of growth).

LH co(n)text	Term	RH co(n)text
all'aumento dei rendimenti seguito alla	**crisi**	finanziaria in Russia. La riduzione dei prestiti
è stata caratterizzata dall'aggravamento della	**crisi**	economica nell'area dell'Est asiatico e nel Giappone
la coesione tra le monete del Sistema. Lo scoppio della	**crisi**	russa in agosto innescava attacchi speculativi,
dell'1,0 per cento (tav. B23). L'eccezionale	**contrazione**	delle esportazioni verso i cinque paesi asiatici più colpiti
e, al suo interno, i membri dell'Uem, hanno registrato una	**contrazione**	del ritmo di crescita a causa di un peggioramento dei saldi
appaiono peraltro in contrasto con i timori di un	**rallentamento**	dell'attività produttiva e con il susseguirsi di fasi
Il contenimento delle spinte salariali, il	**rallentamento**	della congiuntura interna e il cospicuo
Le esportazioni delle regioni italiane Il	**rallentamento**	della crescita delle esportazioni italiane nel 1998
nuovo impulso frenante all'economia mondiale, già colpita dal forte	**rallentamento**	della domanda avvenuto nel corso del 1998.

Table 3b: A concordance of terms related to the business cycle. [Sub-corpus: R (Reports and Speeches)].

A more extensive use of metaphor can be found in newspaper and magazine articles originally written in Italian (cf. Table 3c below): growth is driven by biotechnology (concordance 1), the investment cycle is past its peak (2), budgetary policies could make the cycles or phases of economic slowdown shorter (3), signs of recovery can literally 'feed', i.e. give momentum to the market (4), the business cycle can cool down (4) and US economy is described as the engine that drives the world economy (8). Wide use also leads to processes of derivation as in *decelerazione congiunturale* (business cycle slowdown) and *quadro congiunturale* (business cycle 'picture'). Finally recession is the dreaded 'R' word, which is seen as a scarecrow (8) or fast approaching enemy (9). Here as in the concordances above metaphors are accompanied by evaluative language – strong/weak recovery, moderate growth, slight increase.

LH co(n)text	Term	RH co(n)text
Insomma, sta per iniziare un nuovo	**ciclo economico**	positivo. Una crescita trainata e dominata dalle biotecnologie.
Pur con un utilizzo della capacità produttiva ancora elevato, il	**ciclo degli investimenti**	ha ormai superato il punto di svolta,
limitano il ricorso a politiche di bilancio espansive che accorcerebbero i	**cicli**	di rallentamento economico, e oggi, servirebbero a promuovere
Secondo gli analisti, i segnali di ripresa della	**congiuntura**	internazionale possono dare nuova linfa anche al mercato
mette in evidenza un sensibile peggioramento del	**quadro congiuntu-rale**	a inizio anno, in linea con le indagini qualitative condotte
L'inflazione lontana dagli obiettivi di stabilità e la	**decelerazione congiunturale**	in atto hanno, infatti, spiazzato la Bce, che rinvierà
un aumento dei tassi di interesse per raffreddare la	**congiuntura**	può essere solo graduale. L'alternativa è quella adottata
Eppure il leit motiv è stato lo spauracchio della	**recessione**	Usa. Se la locomotiva non tira che fine farà l'export degli
come gli altri anni. Saranno pochi anche i ricevimenti. Il nemico,	**"R" come recessione,**	non era mai stato così alle porte.

Table 3c: A concordance of terms related to the business cycle. [Sub-corpus: O (Italian originals)].

On the contrary, the language of translated articles is fairly standard, though boom is not only a phase in the business cycles but is also used to mark a surge in the housing market. Interestingly, as many as three synonyms are used in this context – the short form *boom immobiliare*, the extended form *boom del mercato immobiliare*, and *boom dell'edilizia* (cf. Table 3d). In Italian the metaphorical English term boom becomes completely opaque. Its use runs counter to the tendency of avoiding loan words in translation and recognises the dominance of English in the field of economics – and its language.

LH co(n)text	Term	RH co(n)text
non sia mai stato così basso. Alla fine del 2000, il	**boom economico**	ha rallentato assumendo un ritmo più sostenibile,
Allo stesso tempo si assiste in moltissimi paesi ad un	**boom immobiliare,**	dagli Stati Uniti alla Gran Bretagna fino alla Francia
quattro su dieci dei posti di lavoro creati negli ultimi anni sono correlati al	**boom del mercato immobiliare.**	La lezione degli altri paesi che hanno
degli investimenti residen-ziali e un'impen-nata della disoccupa-zione. Grazie al	**boom dell'edilizia,**	gli investimenti resi-denziali ora rappre-sentano il 6%
si verificherà un crollo improvviso, una profonda crisi del dollaro o una forte	**recessione.**	Più presumibilmente si tratterà di una discesa graduale, ma

Table 3d: A concordance of terms related to the business cycle. [Sub-corpus: T (Translated articles)].

Physics can also be used as a source of metaphors for changing economic conditions. The Italian stems *accel** – as in *accelerare* (to accelerate), *acceleratore* (accelerator) etc. –, *decel** (slowdown), and *fren** (brak(e)*) designate rapid or slow economic change. The three physical metaphors identified appear to be more common in the sub-corpora of economic reports and original Italian articles, as shown in Table 4a.

Word/term	Reports+speeches	Orig.Ital.	Transl.Ital.
accel*	**0.2**	**0.2**	0.1
decel*	**0.2**	0.1	0.01
fren*	0.1	**0.5**	0.3

Table 4a: Stems from physics are used to coin terms and describe the pace of economic change.

LH co(n)text	Term	RH co(n)text
Il rallentamento della produttività ha riflesso, in presenza di una moderata	**accelerazione**	del valore aggiunto, la notevole ripresa dell'occupazione
nonostante il calo delle esportazioni del 6,4 per cento, l'attività produttiva è	**accelerata**	rispetto al trimestre precedente, riflettendo il vigore della domanda
nel pubblico impiego, mentre vi è stata una lieve	**accelerazione**	nei settori delle costru-zioni, del commercio e dei servizi
hanno via via compresso la dinamica dei prez-zi, che, dopo la breve	**accelerazione**	registrata all'inizio del 1998, ha toccato minimi storici
Nel 1998 i prestiti in sofferenza hanno fatto registrare una	**decelerazione,**	dal 7,1 al 2,2 per cento (tav. D10);
In un quadro congiunturale carat-terizzato da una lieve	**decelerazione**	dei prezzi al consumo e da una crescita della moneta M3

Table 4b: A concordance of terms from physics [Sub-corpus: R]

The productiveness of the metaphor-like process is also evidenced by extensive derivation and compounding both in the economic report (*accelerazione* and *accelerata* in Table 4b above) and original Italian article sub-corpora (*accelerando*, *frenare*, *freno* and *frenata* in Table 4c below).

LH co(n)text	Term	RH co(n)text
indicare che non ci si è trovati di fronte a una nuova fase di	**accelerazione dei prezzi,**	bensì a impennate "una tantum", di natura tem-poranea,
il trimestre finale dell'anno non mostrerà certo significative	**accelerazioni**	nei ritmi di crescita del Pil, così come della produzione,
lo yuan (renminbi) ha segnato un nuovo record ed è andato	**accelerando**	il passo della rivaluta-zione (si veda il grafico).
L'aumento della produ-zione non servirà a	**frenare**	i prezzi. Perché sul mer-cato c'e già molto greggio.
registrato in entrambi i primi due trimestri, è un importante elemento di	**freno**	dell'attività produttiva. Il deterioramento della do-manda interna
Le esportazioni, a loro volta, sono ripartite dopo la	**frenata**	della scorsa estate e le importazioni hanno mode-ratamente decelerato

Table 4c: A concordance of terms from physics [Sub-corpus: O]

Finally, change in economics is also described by analogy with the weather. A relatively frequent image in Italian is that of *turbolenza/e* (turbulence, disturbance) – especially to describe disturbances in the markets – and *clima* (lit. climate), in particular to term the atmosphere of confidence that is essential for the good performance of the stock exchange. Cold and warm weather as metaphors for the cooling or (over?)heating of the economy is more commonly referred to in translated articles, as frequency of the stems **fred** (cool*) and **cald** (heat*) in Table 5a below shows. Perhaps talk about the weather – either literally or metaphorically – is more central to British than to Italian culture as reflected in their highest frequency in the articles translated from *The Economist*.

Word/term	Reports+speeches	Orig.Ital.	Transl.Ital.
turbolenza/e	**0.1**	0.05	0.01
clima	0.2	**0.3**	0.1
fred	0.01	0.03	**0.2**

Table 5a: The distribution of the tokens related to the fundamental concept of weather per thousand words.

In the sub-corpus of translated articles, the process of cooling appears particularly productive as a KWIC concordance based on the stem **fred** shows in Table 5b. below. Interestingly, this does not seem to have a counterpart in original Italian texts and could thus point to a culture-specific metaphorical process.

LH co(n)text	Term	RH co(n)text
la nuova economia è particolarmente im-portante che i mer-cati recuperino la loro	**freddezza.**	Nel 2000 l'hanno persa. Le valutazioni dei titoli erano tutte alte,
della Cina nel settore dell'acciaio, esacerbato dalle misure adottate per	**raffreddare**	l' economia, potrebbe determinare un'impennata delle
la crescita, la politica fiscale e di spesa potrebbe invece intensificare il	**raffreddamen to economico**	del 2006 con un rischio mortale anche per il futuro.
Anche il mercato		dall'estate scorsa. La bolla

immobiliare britannico	**si è raffreddato**	immobiliare, invece, continua a gonfiarsi.
contemporaneamente della pace nel mondo e della cura per il	**raffreddore.**	In realtà oggi, nessun outsider può pronunciare qualcosa di definitivo

Table 5b: A concordance of terms from the stem *fred* [Sub-corpus: T]

War and conflict as metaphors of economic and business endeavors (Table 6) seem to be more common in translated articles. The stems *lott** (*lotta/e*, fight(s); *lottare*, to fight), *conflitt** (*conflitto/i*, conflict(s), *conflittualità*, relationship(s) based on conflict), and *sfid** (*sfida*, challenge; *sfidare*, to challenge) show lowest frequency in economic reports that mainly deal with macroeconomic topics such as business cycles and overall trends in Italian economy. The metaphors of war and conflict may be better suited to business and finance.

Word/term	Reports+speeches	Orig.Ital.	Transl.Ital.
lott*	-	0.05	**0.3**
conflitt*	0.04	**0.1**	**0.1**
sfid*	0.01	0.1	**0.2**

Table 6. Italian stems used to form metaphors of war and conflict to posit business and finance as a struggle and challenge.

Typically culture-specific metaphors in Italian concern the housing market. In economic reports the general language word *abitazione/i* (house/s) is used or the term for the corresponding market, *mercato immobiliare* (housing market) is referred to. In original Italian articles, however, real estate is the emotionally charged *casa/e* as an icon of family – a key institution in Italian society – and not its economic equivalent, the household. This may also be the case because in the Italian language of economics no distinction is made between family and household, so sentiment can be expressed through *casa* while a neutral attitude is associated with the technical term *mercato immobiliare*. In the language of the press, the building industry and housing market are also metonymically termed by referring to one of their basic components, the *mattone* (brick). As can be seen from Table 7a below, real estate as a culture-specific symbol of family values (*casa*) is not used in economic reports, while both *casa/e* and *immobiliare/i* as an adjective forming compounds relating to housing have very low frequency. *Mattone* – as an instance of Lakoff and Johnson's metaphorical entailments – is not used at all in this sub-corpus.

Word/term	Reports+speeches	Orig.Ital.	Transl.Ital.
casa/e	-	**0.8**	0.01
immobiliare/i	0.03	**0.4**	0.3
mattone	-	**0.2**	-

Table 7a. A powerful metaphor in Italian economics: the *casa* as a symbol of family values.

LH co(n)text	Term	RH co(n)text
Soprattutto nel caso in cui si acquista la prima	**casa,**	che potrebbe non essere quella definitiva, selezionare con cura l'immobile
Dipende da che cosa si vuole. Perché nessuno lascia andare la	**casa**	della vita. Tra gli addetti ai lavori è una sola la parola magica
mettono al primo posto un	**investimento immobiliare**	(39%). Seguono le polizze (21%), le obbligazioni a breve
grazie ai rifinanziamenti, il	**mercato immobiliare**	americano ha messo a segno crescite record e un'ondata di acquisti
alla Commissione bancaria del Senato Usa, Bernanke ha detto che la	**crisi immobiliare**	in corso negli Stati Uniti è per noi un "problema continuo"
Perché la casa non tradisce gli italiani Il	**mattone**	resta d'oro Negli ultimi cinque anni i prezzi degli *immobili* sono saliti anche del 50%
tutti, comprano. Ciascuno è seduto sul suo	**mattone.**	Un primato che rivela. La *casa* è rimasto il principale collante dell'identità nazionale.

Table 7b. A concordance of *casa* and *mattone* as metaphors for family values in Italian [Sub-corpus: O]

An analysis of concordances in the sub-corpus of original Italian articles shows that the three words are used quite differently: while *immobiliare* is involved in the formation of compound terms (*housing investment*, *housing market* and *housing crisis*) and does not include metaphorical language in its co(n)text, when referring to the *casa* but even to the much more technical *mattone*, Italian economic reporters almost wax lyrical about their implications for Italian culture and society. Thus they distinguish between the first 'home' Italians buy (concordance 1) and the emotionally charged 'home of a lifetime' (concordance 2). But even the *mattone* remains the 'gold' investment for Italians (concordance 6) because home as a symbol of family remains the 'glue' (*collante*) that holds together Italian national identity (concordance 7).

On the contrary, the metaphorical concept of bubble (*bolla*) seems to be more current in translated than original Italian texts as it gives rise to a number of compound terms such as *bolla speculative* (speculative bubble), *bolla immobiliare* (housing bubble), *bolla finanziaria* (financial bubble), etc. The term is connoted as it is never used in economic reports, where the neutral *crisi* (crisis) is preferred.

5. Discussion

Analysis in this paper was conducted drawing on a broad concept of metaphor as highlighted by McCloskey (1995) who makes no distinction between metaphors "which draw on similarity, and metonymies, which draw on contiguity or association". Data suggest that use of metaphors varies according to text type. In economic reports metaphors reflect standard use in economics and their surrounding co(n)texts express sentiment in a highly controlled fashion, i.e. with frequent hedging and downtoning. Conversely, newspaper and magazine articles use metaphors more freely by mixing them and

adding expressions typical of a much more emotionally charged language and imagery. Translated texts occupy the middle-ground as they frequently employ what are standard metaphors in the language of economics. In other words, translators prefer to avoid risks and play safe.

In his corpus study of metaphor in business articles, Partington (1998: 107-120) draws on contemporary research on metaphor as a cognitive process, but finally concludes that many metaphors become "genre-specific technical language" and "have no figurative content, and to all intents and purposes are no longer metaphors at all" (1998: 119), so he resorts to the traditional categories of original, standard or cliché, and dead metaphors. While it is true that many metaphors are or become terms, it is more difficult to agree that they lose figurative meaning and are no longer metaphors. This overlooks Kuhn's idea that a metaphor-like process in science continues to play a role until its potential has largely been exploited and even when the metaphor has lost its currency (as outlined in par. 1 above). In the present study an attempt has been made to show how a metaphor-like process is triggered as in the case of *rialzo* → *revisione al rialzo* and the subsequent mixing of metaphors, for example in *rialzare i tassi per gettare acqua sul fuoco dei prezzi e della ripresa*. This is in line with the much debated use of ontological metaphors in science (Ahmad 2006) and with Hundt's idea that metaphors in popular economics are much more extended. In line with Partington's findings, however, the present study found low frequency of metaphors relating to war, no extended metaphors with *flusso* (flow), but frequently mixed metaphors.

Analysis also suggests that some metaphors may be at least partially culture-specific – in line with results of the study of metaphor in general Italian by Deignan and Potter (2004). The clearest case is that of *casa* and *mattone,* but even the higher frequency of metaphors relating to the weather in the translated texts indicate a difference in metaphor-building in Italian and English economics that could be further investigated.

6. Conclusion

In this paper an attempt has been made to show the use of metaphor in shaping one of the key social sciences, economics, by analysing two text types and original and translated texts in Italian. The choice of two text types – reports and newspaper or magazine articles – and the exclusion of academic papers was dictated by the fact that Italian economists mainly write their research papers in English to reach a wider readership and academic papers still written in Italian usually have local interest and can thus make the corpus unbalanced as the topics covered in the reports and articles are mainly macroeconomic or to do with the global economy.

Partly as a result of the multiple methods employed to extract candidate metaphors, the picture of metaphor-building in Italian economics may appear patchy and the number of expressions limited in this paper. The stock of metaphors in the two text types needs to be added to. For this reason, the corpus is currently being expanded to include original and translated articles from other Italian newspapers and magazines and the economic reports and speeches sub-corpus is being updated by including more reports published by the Bank of Italy and ISTAT, and speeches delivered by the current Governor of the Bank of Italy. It will then be possible to provide a better picture of universal or culture-specific metaphors, and to extend investigation to other figures of speech, their collocation patterns and their co(n)texts.

7. References

Ahmad, K. (2006). Metaphors in the languages of science? (pp. 197-220). In: Gotti, M. & Bhatia, V. (eds.) *New Trends in Specialized Discourse Analysis*. Bern: Peter Lang,.

Black, M. (1962). *Models and Metaphors: Studies in Language and Philosophy*. Ithaca, NY: Cornell University Press.

Boyd, R. (1979). Metaphor and theory change: What is "metaphor" a metaphor for? (pp. 356-408). In Ortony, A. (ed.) *Metaphor and Thought*. Cambridge: Cambridge University Press.

Cruse, A. (2004) *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Deignan, A. & Potter, L. (2004). A corpus study of metaphors and metonyms in English and Italian (pp. 1231-1352). In: *Journal of Pragmatics*, 36.

Eubanks, P. (1999). Conceptual metaphor as rhetorical response. A reconsideration of metaphor (pp. 171-199). In: *Written Communication*, 16 (2).

Henderson, W. (1982). Metaphor in economics (pp. 147-157) *Economics*, 18 (4).

Henderson, W. (1993). The problem of Edgeworth's style (pp. 200-222). In: Henderson, W., Dudley-Evans, T. & Backhouse, R. (1993). *Economics and Language*. London: Routledge.

Hundt M. (1998). Typologien der Wirtschaftssprache: Spekulation oder Notwendigkeit? (pp. 98-115). In: *Fachsprache*, 20 (3-4).

Klamer, A. & Leonard, T.C. (1993). So what's an economic metaphor? (pp. 20-53). In: Mirowski, P. (ed.) *Natural Images in Economic Thought*. Cambridge: Cambridge University Press.

Kövecses, Z. (2002). *Metaphor. A Practical Introduction*. Oxford/New York: Oxford University Press.

Kuhn, T. (1979). Metaphor in science (pp. 409-419). In Ortony A. (ed.) *Metaphor and Thought*. Cambridge: Cambridge University Press.

Lakoff, G. & Johnson, M. (1980/2003). *Metaphors We Live By*. Chicago: University of Chicago Press.

McCloskey D. (1988). *La retorica dell'economia* (Italian translation of *The Rhetoric of Economics*). Torino: Einaudi.

McCloskey, D. (1995). Metaphors economists live by. *Social Research*, Summer, http://findarticles.com/p/articles/mi_m2267/is_n2_v62/ai_17464378 (last visited on March 1, 2007).

Partington, A. (1998) *Patterns and Meanings. Using Corpora in English Language Research and Teaching.* Amsterdam/Philadelphia: John Benjamins.

Resche, C. (2000) Equivocal economic terms or terminology revisited (pp. 158-173). In *Meta*, 45 (1).

Resche, C. (2004) Approche “terminométrique” du cicle économique: implications et prolongements (pp. 343-359). In *Meta*, 49 (2).

Richards, I.A. (1965). *The Philosophy of Rhetoric*. New York: Oxford University Press.

Sperber, D. & Wilson, D. (1986/1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Now you see them, now you don't: Lexicalised metaphors in translation

Margaret Rogers,
Centre for Translation Studies,
University of Surrey, Guildford, Surrey, UK. GU2 5XH
m.rogers@surrey.ac.uk

Abstract (only)

Early work within the relatively new discipline of Translation Studies suggested that metaphorically derived polysemous senses of codified words should no longer be considered as metaphors (Dagut 1976). Words such as 'foot' (of the mountain), 'arm' (of an organisation), 'leg' (support for material and immaterial objects in general), 'head' (on beer), and so on, would not, under this understanding, be considered to be metaphorical. But if such figurative senses are excluded from an analysis of metaphorical patternings, then important insights into the ways in which different languages map objects (both material and immaterial) may be lost. It is well known, for instance, that translation equivalents–here understood as dictionary equivalents–do not necessarily reflect the metaphorisation patterns of other languages. So in German, for instance, literal:figurative lexical patterns are different from those in English: 'arm' is *Arm* (literal: body part) or *Ast* (figurative: part of an organisation; literally 'branch'), 'head' is *Kopf* (literal: body part) or *Blume* (figurative: foamy layer on beer; literally 'flower'). This suggests that the underlying conceptual metaphors (Lakoff & Johnson 1980) by which different cultures linguistically map aspects of the world are also different.

One of the differences claimed to distinguish terms, as the linguistic expression of knowledge-rich specialised concepts, from general-purpose lexical items or words, is that terms are purely denotative and words both denotative and connotative (cf. for example, Felber 1984:98 for a categorical view). Other views emphasise the importance of metaphorisation as a means of facilitating understanding of evolving domains in which '[t]he proof and results of metaphorical thinking are in the metaphorical lexicalisations' (Temmerman 2000:156). These very different views reveal philosophically contrasting approaches: on the one hand, terms are seen as linguistic designations of an objective reality, while on the other hand, terms are seen as the outcome of experiential human attempts to understand and order the world. In this paper, it will be argued that a translational perspective on metaphor—including equivalence patterns of codified words and terms—can potentially reveal alternative conceptions of domains in the lexical patternings found in special-language texts. Illustrations will be drawn from parallel texts in the field of economics, focusing on German and English.

References

Dagut, M. (1976). Can metaphor be translated? *Babel: International Journal of Translation*, *XXII* (1), 21-33.

Felber, H. (1984). *Terminology Manual*. Paris: UNESCO/Infoterm.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Temmerman, R. (2000). *Towards New Ways of Terminology Description. The sociocognitive approach*. Amsterdam/Philadelphia: Benjamins.

Applications

Detecting Uncertainty in Spoken Dialogues: An explorative research to the automatic detection of a speakers' uncertainty by using prosodic markers

Jeroen Dral, Dirk Heylen and Rieks op den Akker
University of Twente
the Netherlands
j.s.dral@student.utwente.nl, heylen@cs.utwente.nl, infrieks@cs.utwente.nl

Abstract

This paper reports results in automatic detection of speakers uncertainty in spoken dialogues by using prosodic markers. For this purpose a substantial part of the AMI corpus (a multi-modal multi-party meeting corpus) has been selected and converted to a suitable format so its data could be analyzed for selected prosodic features. In the absence of relevant stance annotations on (un)certainty, lexical markers (hedges) have been used to mark utterances as either certain, or uncertain. Results show that prosodic features can indeed be used to detect speaker uncertainty in spoken dialogues. The classifiers can distinguish uncertain from neutral utterances with an accuracy of 75% which is 25% over the baseline.

1. Introduction

Each utterance we make comes with a particular degree of certainty we have about the state of affairs that is described in our utterance actually holding. We may feel reasonably confident or rather hesitant about whether there is any truth in what we are saying. We often express this degree of certainty in what we are saying through hedges ("I think"), modal verbs ("might"), adverbs ("probably"), tone of voice, intonation, hesitations. We can accompany the speech with gestures and facial expressions that can express the same hesitant or confident state of mind. This research will focus on the prosodic features of speech and will try to develop a method to automatically classify speech as being (un)certain. The purpose of this research is to (automatically) measure one's belief (or confidence or self-conviction) in the correctness of a certain utterance. Even when the definition of uncertainty is clear, the question remains how to state the degree of uncertainty? Is it certain or uncertain or are there shades of gray in between? And if so, how do we state them?

2. Related Works

Although uncertainty can be detected by both visual and non-visual means this research, and the overview of the related work, will focus on the non-visual aspects of the detection of (un)certainty.

2. Defining (un)certainty

People's ability to accurately assess and monitor their own knowledge has been called the 'feeling of knowing' or FOK by Hart [1]. Many experiments on this area are based on question-answering where respondents must answer certain (knowledge) questions and assess whether their answer is likely to be correct. A study by Smith and Clark [2] investigated FOK in a conversational setting and followed the method mentioned above. Respondents were asked to answer general knowledge questions, then estimated their FOK about these questions and finally were tested on their ability to recognize the correct answer. They found that FOK was positively correlated with recognition and with response latency when retrieval failed and negatively correlated when retrieval succeeded. Another study by Brennan and Williams [3] used the research of Smith and Clark and in addition researched the sensitivity of listeners to the intonation of answers, latencies to responses and the form of non-answers. When looking at the 'feeling of another's knowing' or FOAK, Brennan and Williams state a listener can use several different sources of information to evaluate a respondent's knowledge:

- His own knowledge
- Assess the difficulty of the question for the average person or for the typical member of a particular community and use that information to judge a respondent's confidence.
- Information from their shared physical environment and from immediately previous conversation ("mutual knowledge").
- Information about the respondent's ability or previous performance
- Paralinguistic information displayed in the surface features of respondent's responses (intonation, latency to response).

In their experiments they concentrated on the paralinguistic information available. The result of their experiments supports the interactive model of question-answering and shows the display of

respondent's metacognitive states when searching their memories for an answer. Another conclusion which can be made based on their research is the ability of listeners to use these cues. Their FOAK was affected by the intonation of answers, the form of non-answers and the latency to response (e.g. a rising intonation often accompanied a wrong answer).

Krahmer and Swerts[4] describe experiments with adults and children on signaling and detecting of uncertainty in audiovisual speech. They found that when adults feel uncertain about their answer they more likely produce filled pauses, delays and higher intonation (as well as some visual signals, such as eyebrow movements, and smiles). For child speakers similar results are measured but less prominent. The children in this experiment were aged 7-8, which is younger than the children in Rowlands study[5], who were over 10. (see next subsection). Age matters: Krahmer and Swerts suggest that young children do not signal uncertainty in the way adults and older children do because they care less about self-presentation than adults. Our study is about adult subjects only.

2.2 Linguistic pointers to uncertainty

Knowledge questions can be seen as 'testing questions' where the focus may not be on revealing the truth but rather on exposing ignorance and thus adding pressure to the speaker, making him nervous and uncertain; see Ainley's study[6]. Since a common perception about mathematical propositions is that they are either right or wrong, Rowland analyses transcripts of interviews with children focused on mathematical tasks and looks at the children's use of language to shield themselves against accusation of error [5]. According to his research children tend to use a certain category of words (called *hedges*) which are associated with uncertainty. These hedges are further divided in different types:

- Shield
    - Plausibility shield (I think, maybe, probably)
    - Attribution shield (According to, says...)
- Approximators
    - Rounders (About, around, approximately)
    - Adaptor (A little bit, somewhat, fairly)

While some hedges are obvious shields to 'failure' others are more elusive and require some contextual information. For example, the word 'about' may be a shield when used in combination with a number (e.g. 'there live about 150 thousand people in Enschede') but is no such thing when used in a sentence like 'the story is about a small boy'.

Another research which looks at the use of hedges is that of Bhatt et al [7]. In their research they study how students hedge and express affect when interacting with both humans and computer systems. It was found that students hedge and apologize to human tutors often, but very rarely to computer tutors. Another important result of their research is that hedging is not a clear indicator of student uncertainty or misunderstanding, but rather connected to issues of conversational flow and politeness.

2.3 Prosodic markers of uncertainty

Prosody is important because a speaker can communicate different meanings not extractable from lexical cues by giving acoustic 'instructions' to the listener how to interpret the speech. A good example is the increasing pitch (high F_0) at the end of a question. By using this kind of intonation the speaker draws attention to his question. Other theories include the speaker taking a humble stance by imitating a younger person (with higher F_0 and formants) since he's actually asking a favour to the listener (answering his question) [8, pp. 277].

In their research Liscombe et al. investigate the role of affect (student certainness) in spoken tutorial systems and whether it is automatically detectable by using prosody [9]. They discovered that tutors respond differently to uncertain students than to certain ones. Experiments with Intelligent Tutorial Systems (ITS) indicate that it is also possible to automatically detect student uncertainness and utilize that knowledge for improvement of these ITS's, making them more humanlike. During their research they not only looked at the current (speaker) turn but also compared this turn with the dialogue history. Among the features analyzed were mean, minimum, maximum and standard deviation statistics of F_0 and the intensity, voiced frames ratios, turn duration and relative positions where certain events occurred.

3. PROBLEM STATEMENT

Since much of the research above limits itself to the answering of trivia questions or short answers some question marks can be placed at the usefulness of the results in a broader/different context. Many applications using automatic recognition of the degree of certainty of a person with respect to what he is saying might require different input than 'simple' question/answer-pairs. Since the experiments as described above needed relatively

short answers (a few words) in order to get a standardized intonation (see [10]) one could wonder what the effects will be on longer utterances like normal dialogues, statements or presentations. Also, a rising intonation (a sign of uncertainty when answering a question) then can also be meant as a question itself (so how to differentiate between the two?) and the latency before an utterance may be irrelevant since the (potential) uncertain utterance might be encapsulated in other utterances from the same speaker. Nonetheless, these short utterances derived from question answering sessions make it possible to research prosodic features of speech which may be correlated with (un)certainty.

Can prosodic features be used to automatically assess the degree of (un)certainty in a normal spoken dialog? And which features, if any, qualify best as prosodic markers to the qualification of this (un)certainty?

From previous research we already saw that certain features (intonation, latency) can be used to assess the degree of (un)certainty in (short) answers to questions. While the applicability of these features on utterance derived from normal dialogue may be a bit more complex they are still expected to be valuable indicators. Uncertain utterances will probably have a rising intonation due to the questionable nature of these utterances ("Maybe we can make a green remote?"). Also, common sense would correlate uncertain utterances with longer pauses (latencies) between words.

Besides intonation and latency (or gaps between words in case of longer utterances) I can imagine intensity (softer, less conviction in case of uncertainty) and the speed of talking to be a factor to identify uncertainty. In both cases some way of comparing it to a mean value for these features will be needed though since it wouldn't be possible to state whether the utterance has a below/above average value for intensity or speed.

4. DATA SELECTION

In order to be able to perform prosodic analysis and reach some valid conclusions, it seemed logical to use an existing corpus which had already been annotated. The AMI Corpus[11], which we addressed during the preliminary phase of this project, not only had many hours of high quality voice recordings but also annotations on different levels (hand made speech transcriptions, time aligned words, dialog acts) which could be used for this research.

4.1 Selection of Meetings

After reviewing the available annotation data for the AMI Corpus (public release 1.3.1) a choice had to be made as to which sets were to be analyzed. Since the ES, IS and TS sets were the only ones with complete coverage of the words and dialog acts annotations and the existence of these annotations was considered essential these three sets were chosen. As can be seen in 1 the total dataset now existed of 552 audio files with a total duration of about 280 hours.

Table 1 Overview of selected audio files

	Groups	Meetings	Files	Duration
ES	15	60	240	118:52:35
IS	10	40	152	93:05:28
TS	10	40	160	92:54:05
Total	35	140	552	278:01:50

A disadvantage of the corpus used is the lack of sufficient stance annotations needed for the identification of uncertainty in speech. Since there was no reliable and efficient way to mark uncertain utterances, it was decided to use lexical elements (hedges) to identify utterances which would have a high probability of being uncertain. We split the dialogue acts into three classes: *uncertain* (that contain uncertainty hedges), *certain* (that contain certain hedges), and *neutral* (that do not contain any hedges).

Table 2. Overview of hedges for uncertainty and words indicating certainty

Uncertainty	Certainty
according (to)	absolutely
approximately	certainly
around	clearly
fairly	definitely
maybe	(in) fact
perhaps	must
possible	obviously
possibly	(of) course
probable	positively
probably	surely
somewhat	undeniably
(I) think	undoubtedly
usually	

In Table 2 an overview of indicators used can be seen. These groups of words are derived from previous studies as performed by Rowland [5] and Bhatt et al [7]. This approach raises some questions. In their study Bhatt et al already disputed hedges

being only indicators for uncertainty, mentioning they could also be used for politeness strategies [7]. To make sure the assumption made was valid 25 random dialog acts, marked as uncertain during this research, were ranked on a five point scale ranging from certain to uncertain: certain- probably certain – undecided – probably uncertain – uncertain. 80% of the utterances were scored as either uncertain or probably uncertain.

4.2 Data Preparation and Selection

In preparing the AMI data to run through PRAAT[12], certain errors in the data were found (missing end or begin times of words). Since the Dialog Act tiers are based on the word tiers therefore several Dialog Act intervals had missing start and/or end times also and had to be discarded. In Table 3 the total amount of valid and invalid items can be seen. Since the percentage of these incorrectly annotated words and dialog acts was very low it was decided to simply discard them from the dataset instead of trying to figure out the correct data (if possible at all).

Table 3 Overview of converted words and dialog acts

	Words			Dialog Acts		
Series	Valid	Invalid	Invalid %	Valid	Invalid	Invalid %
ES	351.615	42	0,01%	47.251	35	0,07%
IS	198.968	14	0,01%	26.909	14	0,05%
TS	283.208	695	0,24%	42.394	419	0,98%
Tot	833.791	751	0,09%	116.554	468	0,40%

We used PRAAT for prosodic analysis. First a selection of the relevant prosodic features was made.

For each category of the prosodic properties mentioned in section 2.3, several attributes were chosen and implemented in PRAAT. Beside these prosodic attributes some lexical attributes (like the number of words, the presence of 'yeah (, but)', 'okay') were added as well. In total 76 attributes were chosen for the analysis, of which 67 were prosodic.

The amount of dialog acts including hedges consists of only 7,26% of the total (7.317 dialog acts of a total of 100.799), which means that simply classifying each dialog act as certain gives a score of about 93%. By balancing the dataset the script will take 4.819 random other dialog acts and combine them with the ones containing hedges to form a new dataset.

4.3 Statistical Analysis

Since the dataset preparation script in phase 4 has been designed in such a way that different datasets can be created on the fly it is easy to compare different prosodic features of different classes. In phase 3 of the research, the actual prosodic analysis, the presence of several lexical markers or indicators was also checked. Among these markers were the hedges as mentioned before, the group of words (supposedly) indicating certainty, yeah and okay.

5. EXPERIMENTATION

During the following experiments all datasets were leveled on a 50/50 basis so each 'group' was equally represented. As a result the baseline (computed with the ZeroR classifier) of all datasets is about 50%. Next, the datasets were classified with the J48 (tree) and NaiveBayes (NB) classifiers. Each classifier was evaluated for accuracy using 10-fold cross-validation. We used the implementation in the Weka toolkit[13]. To determine the key attributes being used for this classification the input data was also evaluated using the InfoGain attribute evaluator in combination with a Ranker search method.

5.1 Uncertain Hedges –vs– Neutral

First the dataset with the hedges was analyzed. Out of all 100.799 dialog acts analyzed with PRAAT in phase 3 only 7.317 contained one or more hedges (see also 4). These instances were complemented with the same (random) amount of dialog acts containing no hedges; the neutral set. Based on previous research it was expected that several prosodic features would be good indicators for uncertainty in speech. Among these features were a rising pitch, a declining intensity and a slower rate of speech (more pauses and/or longer average word-length).

Table 4. Properties of dataset Uncertain Hedges –vs– No Hedges

Class	Instances
No Hedges (neutral)	7.317 (dropped 85.502)
Uncertain Hedges	7.317

In Table 5 the results of the analysis can be seen. Two classifiers were used (J48 and NaiveBayes); for each the improvement over the baseline (IOB) is included in the table. As anticipated the baseline is about 50% correct classifications. Two striking results are the overall improvement over the baseline score (with an average increase of about

17/18% based on which classifier has been used) and the high performance on the lexical features alone. The evaluation of the (key) attributes show the importance of attributes related to the length of the dialog act.

Table 5. Classification Performance of Hedges –vs– No Hedges including improvement over baseline (IOB)

Baseline (ZeroR)	49,98%			
Features	**J48**	**IOB**	**NB**	**IOB**
Lexical features (LF)	74,67%	24,69%	71,27%	21,29%
Spectrum related features (SF)	67,70%	17,72%	64,59%	14,61%
Pitch related features (PF)	68,27%	18,29%	67,04%	17,06%
Intensity related features (IF)	63,61%	13,63%	61,20%	11,22%
Formant related features (FF)	66,80%	16,82%	67,97%	17,99%
All Prosodic Features	66,05%	16,07%	68,46%	18,48%
All features	71,14%	21,16%	69,96%	19,98%
Average Improvement		18,34%		17,23%

The first 8 attributes, headed by the amount of words (da_words) in the DA, are all related to the DA length, either indicating time or the amount of (voiced) frames or bins. Since the utterances in the corpus have been marked (un)certain by using hedges this is not very surprising: hedges are normally part of (longer) sentences. As a result, the length of a dialog act (shown by a number of attributes) is a good indicator since short dialog acts are often marked certain.

After the attributes indicating length in some way the type of dialog act is also important, taking a 9th place in the attribute ranking. Apparently the type of DA as annotated by the members of the AMI Project has some relation to uncertainty. More about the distribution of hedges over dialog acts can be seen in section 5.3.

Next in the attribute ranking are several formant attributes headed by the minimum F_2, maximum F_1 and maximum F_2. After several other formant attributes the standard deviation for the intensity during the 2nd half of the DA, the spectrum band energy, and the voiced frame ration during the 2nd half, and the total DA seem to be good indicators for uncertainty. When classifying the dataset with the J48 classifier and using only the formants' minimum and maximum values the performance result is 67,3%, even higher than when using all formant attributes. Classification based on the voiced frame ratios only gives a performance of 59,8%.

5.2 Uncertain Hedges –vs– Certain Hedges

Similar to the previous dataset where dialog acts with hedges were compared to dialog acts without these lexical markers another set was created which contained all dialog acts with words which should indicate certainty and compared to a similar sized group of hedged dialog acts. As can be seen in 6 the size of this dataset was significantly smaller.

Table 6. Properties of dataset Uncertain Hedges –vs– Certain Hedges

Class	Instances
UncertainHedges	663 (dropped 6.654)
Certain Hedges	663

Table 7. Classification Performance of Uncertain Hedges –vs– Certain Hedges including improvement over baseline (IOB)

Baseline (ZeroR)	49,77%			
Features	**J48**	**IOB**	**NB**	**IOB**
Lexical features (LF)	58,30%	8,52%	57,77%	7,99%
Spectrum related features (SF)	55,66%	5,88%	50,38%	0,60%
Pitch related features (PF)	55,13%	5,35%	52,26%	2,49%
Intensity related features (IF)	53,09%	3,32%	54,45%	4,68%
Formant related features (FF)	51,58%	1,81%	55,13%	5,35%
All Prosodic Features	55,28%	5,51%	54,90%	5,13%
All features	56,41%	6,64%	55,51%	5,73%
Average Improvement		5,29%		4,57%

In contrast with the expectations mentioned above the actual results show a lower performance of the classifiers with an average improvement of about 5%. Once again the lexical features score best, although the gap is smaller.

This time the attribute ranking shows the type of DA (da_type) being the most predictive attribute, followed by some length related attributes. The first prosodic feature is the mean F_4 (6th place), followed by the minimum intensity (9th) and minimum pitch (12th). In contrast to the previous dataset where the formants played an important role, for this dataset the pitch values (mainly of the 2nd half of the DA) seem to be a better indicator for uncertainty.

5.3 Distribution of hedges over dialog acts

To see whether uncertain utterances occur more in particular dialog acts the distribution of dialog acts marked uncertain over the different dialog act classes has been looked into, the results of which can be seen in Table 8. For comparison, the distribution of all dialog acts has been included as well.

Table 8. Distribution of (uncertain) dialog acts

Dialog Acts (ID)	Total Dialog Acts	Percentage of Total DA's	Hedges	Percentage of Hedges	Percentage of Dialog Act
Minor	30.816	30,6%	670	9,2%	2,2%
Backchannel (1)	10.655	10,6%	33	0,5%	0,3%
Stall (2)	6.983	6,9%	82	1,1%	1,2%
Fragment (3)	13.178	13,1%	555	7,6%	4,2%
Task	56.438	56,0%	6.094	83,3%	10,8%
Inform (4)	29.841	29,6%	2.456	33,6%	8,2%
Suggest (6)	8.610	8,5%	1.645	22,5%	19,1%
Assess (9)	17.987	17,8%	1.993	27,2%	11,1%
Elicit	6.557	6,5%	396	5,4%	6,0%
Elicit-Inform (5)	3.743	3,7%	125	1,7%	3,3%
Elicit-Offer-Or-Suggest (8)	640	0,6%	45	0,6%	7,0%
Elicit-Assessment (11)	2.016	2,0%	225	3,1%	11,2%
Elicit-Comment-Understanding (13)	158	0,2%	1	0,0%	0,6%
Other	6.988	6,9%	157	2,1%	2,2%
Offer (7)	1.370	1,4%	80	1,1%	5,8%
Comment-About-Understanding (12)	1.942	1,9%	16	0,2%	0,8%
Be-Positive (14)	1.856	1,8%	40	0,5%	2,2%
Be-Negative (15)	84	0,1%	3	0,0%	3,6%
Other (16)	1.736	1,7%	18	0,2%	1,0%
Total	100.799	100,0%	7.317	100,0%	5,5%

As can be seen in Table 8 most dialog acts are task oriented or minor acts (56% and 31% respectively). We can also notice that most dialog acts marked as uncertain (i.e. containing uncertain hedges) belong to the task-category.

The class of minor acts contains significantly less uncertain hedges than the class of elicit acts ($Chi^2(df=1)=311.45$; $p<0.001$) and the class of elicits contains significantly less of these hedges than the class of task acts ($Chi^2(df=1)=143.93$; $p<0.001$).

6. CONCLUSIONS

Based on the results described in the previous paragraphs, with classification performance increases of up to more than 20%, it is feasible to conclude that the degree of (un)certainty in spoken dialogues can be assessed automatically. When looking at the features which qualify best as prosodic markers to uncertainty the textual features obviously score best. Due to the nature of the uncertain utterances (being based on hedges which most often require some sort of sentence) this result might be of no surprise. There also seems to be a connection between the type of dialog acts (as annotated by members of the AMI Project) and the degree of uncertainty since the presence of uncertain utterances in several dialog act types is clearly above average. A relatively high percentage of uncertain dialog acts are suggestions or assessments. Whether these dialog acts are really uncertain or whether politeness strategies play a role here is hard to establish.

Another interesting point are the results on which prosodic markers qualify best. It was predicted that a rising intonation, longer pauses (latencies) and a decreasing intensity would be good indicators for uncertainty. Based on the attribute evaluation of the different datasets these theories seem to be supported, showing important roles for the pitch and intensity features. Especially with the dataset 'Hedges –vs– No Hedges' the minimum and maximum values of the formants are good prosodic markers as well.

Even though the results seem straightforward, with impressive classifier improvements over the baseline performances, several questions still remain.

In the current research the feature extraction was based on previous research and the possibilities of PRAAT. While a broad range of features have been researched it could very well be certain additional features might be promising as well. Another improvement could be using custom settings in PRAAT. For now all settings have been kept on default but it is known that, for optimal results, different settings should be used for men and women for example. Additional difficulty would be to either automatically detect the gender of a speaker and adapt the settings accordingly, or manually set gender-values for all 500+ files.

For future research on this topic it would be advisable to have a clear understanding of what the 'uncertainty' being researched entitles and how it can be measured. Having that information should provide a basis for reliable annotations, with which further research can be done. Further research in hedges and/or other lexical markers as indicators for uncertainty looks promising. The results of combined feature sets already showed the best results and expanding those features with other indicators (also visual) will probably give the best results in the end (although not all types of information will be available in all situations).

REFERENCES

[1] J. T. Hart, "Memory and the feeling-of-knowing experience," *Journal of Educational Psychology,* vol. 56, pp. 208-216, 1965.

[2] V. L. Smith and H. H. Clark, "On the Course of Answering Questions," *Journal of Memory and Language,* vol. 32, pp. 25-38, Feb 1993.

[3] S. E. Brennan and M. Williams, "The feeling of anothers knowing - prosody and filled pauses as clues to listeners about the metacognitive states of speakers," *Journal of Memory and Language,* vol. 34, pp. 383-398, Jun 1995.

[4] E. Krahmer and M. Swerts, How children and adults signal and detect uncertainty in audovisual speech. Language and Speech, 48(1):29-54.

[5] T. Rowland, "Hedges in mathematics talk: Linguistic pointers to uncertainty," *Educational Studies in Mathematics,* vol. 29, pp. 327-353, December 1995.

[6] J. Ainley, "Perceptions of teachers' questioning styles," in *12th International Conference for the Psychology of Mathematics Education*, Vezprém, Hungary, 1988, pp. 92-99.

[7] K. Bhatt, M. Evens, and S. Argamon, "Hedged Responses and Expressions of Affect in Human/Human and Human/Computer Tutorial Interactions," in *26th annual meeting of the Cognitive Science Society*, Chicago, Illinois, USA, 2004.

[8] A. C. M. Rietveld and V. J. Van Heuven, *Algemene Fonetiek*. Bussum: Uitgeverij Coutinho, 1997.

[9] J. Liscombe, J. Hirschberg, and J. J. Venditti, "Detecting Certainness in Spoken Tutorial Dialogues," in *9th European Conference on Speech Communication and Technology*, Lisbon, 2005, pp. 1837-1840.

[10] Y. Ozuru and W. Hirst, "Surface features of utterances, credibility judgments, and memory," *Memory & Cognition,* vol. 34, pp. 1512-1526, Oct 2006.

[11] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M.Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meeting corpus. In Measuring Behaviour, Proceedings of 5th International Conference on Methods and Techniques in Behavioral Research.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," January 2008, version 5.0.06 [Software]. Available: http://www.praat.org/.

[13] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2000

POSTER

Motion and Emotion or how to align emotional cues with game actions

G. Lortal[1], C. Mathon[2]
[1]Thales Research & Technology
Route départementale 128 - 91767 Palaiseau cedex – France
[2]EA 333 ARP, UFRL, Université Paris Diderot Paris 7
Case 7003, 30 rue du Château des rentiers, 75205 Paris CEDEX 13- France
E-mail: gaelle.lortal@thalesgroup.com, mathon@linguist.jussieu.fr

Abstract

This paper reports a preliminary study on expression of emotion in sport in French. The study is conducted on texts and natural speech data, recorded from comments of sportive events during the Rugby World Cup 2007. Our aim is to semi-automatically extract passages representative from emotional behaviours in sports comments. As we consider that it is necessary to combine several viewpoints on a same object to avoid inconsistencies, we collected a set of various types of data about one sequence. For one video match, we recorded audio and textual data from professional and non-professional speakers. To extract representative passages, we align an emotional profile defined by linguistic analysis (i.e. prosodic and terminological parameters) with sequences of a sport action. The sequences are retrieved as segments of signal with their transcription. The alignment is made by the matching of a termino-ontological resource with several viewpoints (emotion, sport, discourse types) of several granularities (prosody, emotional terms and sports lexicon). It will be used to permit automatic emotion segments extraction in multimedia documents for corpus collection.

1. Introduction

The emotion of winning or emotional becoming of sportsmen and women is well known. But supporters also ride on high emotion to victory. The emotion that is interesting us here is the emotion that makes supporters "cheering" but also "praying" for the success of their team as it was said in newspapers during the Rugby World Cup 2007. Without trying to understand why sport is leading to emotion, we consider that the behaviour of supporter watching a match is well representative of emotion. We also consider that supporters become all the more emotional when some crucial actions in the match happen.

We consider emotion as defined by K.R. Scherer (Scherer, 2000). In the cognitivist approach, emotion is a relatively brief phenomenon, a reply to the organism according to the evaluation of an external or internal event.

Our aim is to extract passages representative from emotional behaviours in a corpus of Rugby World Cup comments.

We want to align emotional profile (a set of prosodic and terminological parameters manually defined for active emotion by an analysis) with segments of signal representative from a sport action. The alignment is made by the matching of parts of ontologies with several viewpoints (emotion, sport, discourse types) of several granularities (prosody, emotional terms and sports lexicon).

In this paper, we first present our positioning on corpus and emotion. We then describe the corpus we used and how it has been collected. We then expose how the Termino-Ontological viewpoint of the sport ontology has been built. The fourth part explains the prosodic analysis led on a collection of manually extracted segments of sport action and providing us with an active emotion profile. Lastly, we explain how we use the built ontology to index segments of signal by means of a corpus analysis tool enabling lexical, prosodic and discourse analysis.

2. Positioning

In the emotion field and especially in the numerous studies about vocal expression of emotions, the corpus is a central problem. Its choice is not trivial.

A lot of research works are based on simulated emotional expression, with actors (professional or not) playing roles (Fónagy, 1983), (Léon, 1993), (Scherer, 1995). This allows for a tighter control of the quality of the recordings and also to select the emotion to be acted.

Moreover, recently more and more studies have insisted on the necessity of using natural emotional speech. It enables to access authentic expressions (Douglas-Cowie *et al.*, 2000).

However, this latter choice raises difficulties. To obtain this data brings up ethical as well as practical difficulties. It's hard to have good sound quality records outside anechoic room (Campbell, 2001).

The media form an interesting source for spontaneous speech (Chung, 2000) but it could be harder to find a discourse type favouring emotion expression.

But sport is a phenomenon likely to arouse emotion. We all know the jubilant or desperate crowd of supporters after a final world cup match.

In addition to that, a lot of sports meeting are broadcasted, commented and are the subject of journalistic papers. These comments, oral or written, often provide authentic emotional expressions.

The idea here is to "duplicate" our sports emotion oral corpus with written commentaries in order to deal with prosodic parameters, semantic parameters and pragmatic parameters.

To enable a pragmatic analysis of the emotion in sports, we foresee two categories of commentators. The first types of comments are produced by journalists or official commentators and the second by supporters behind the TV.

We claim that our corpus enables us to tackle the

question of expression of emotion from different linguistic approaches, as listed in (Kerbrat-Orecchioni, 2000): lexical approach, morpho-syntactic approach, pragmatic approach, interactionnist approach including the intercultural variation.
Likewise, it helps us considering the nature of such emotional expression, as sport emotion is not straightforwardly listed as basic emotion (anger, joy, sadness, fear, disgust).
Now we present the corpus (oral and written) we use to define emotion parameters within our context.

3. Corpus Description

3.1 Oral

We chose to work on a spontaneous speech corpus and specifically on a sports comments collection recorded during the Rugby Word Cup 2007. The collect was carried out from September the 7th to October the 20th.
Our oral corpus is formed of two sub-corpora corresponding to two types of speaker and record.

3.1.1. Journalistic corpus

Our first sub-corpus consists of a sports comments expressed by journalist. They comment the live-broadcasted matches of the championship.
The matches have been digitized in order to keep the audio sound, mono, with a sampling frequency of 22050 Hz or 44100 Hz.
We recorded the comments in several languages, from the French, Italian, English and Japanese TV.
We collected the broadcasting of 19 matches for French, which is approximately equivalent to 36 hours of recording. For other languages, the corpora are lighter, but we want to process them differently in order to compare cultural behaviours. We recorded only two matches / 3.3 hours for English, 3 matches / 5.6 hours for Japanese, and one match / 1.9 hour for Italian. The journalistic oral corpus is then formed by 25 matches or a 46.8 hours record.
The comment is shared by several speakers. Usually, the action is described by the journalist who is the main commentator of the match. That's why he has the longest speech duration. A retired player is associated as a specialist. He permits to create an interaction with the journalist. When a critical phase of the game is playing or has been played, he is explaining some strategic elements often as a dialog with the commentator. This dialog aims at explaining an action to the viewer audience in a more vivid way.
A third speaker sometimes appeared. He is near the pitch and intervenes when some changes are made in the composition of the team or some exchanges occurred on the field among the players and the referee for example.

3.1.2. Supporter corpus

The 2nd sub-corpus is different from the journalistic corpus by the type of speakers and of record.
We recorded the comments of supporters watching the game at the TV or on an open-air screen. This corpus is only in French.
We collected sound recording for 18 speakers, 14 men and 4 women, during 7 different matches, which represents 10.55 hours of recording.
This corpus is characterized by its amount of overlapping and "sounds". More than 200 "noises" of different types are recorded: breathing, exhalation, puff, sounds of the mouth, sounds of the throat, cough, laugh, sneer, whistle and hiss…

	Oral	Written
Professional comment	1.9 hrs*	31 000 words
Supporter comments	1.5 hrs*	4000 words
*Cultural/languages and sex comparison enabled		

Table 1: Average corpus size for one match

The quantitative description of the corpus is given by match because all data is not yet available for the written corpus. In fact, the written corpus is transcription of the oral one with the tool Transcriber [1]. The transcription is a time-costly activity and is not finished yet for the 36 hours of recording in French. Likewise, the duration of recording is not representative of the speech duration itself.

3.2 Written

3.2.1. Journalistic corpus

The journalistic written sub-corpus is collected from two main sources. The first one is newspapers - mainly on-line - (20 minutes, sport365, Le Figaro, sport24, afp, lexpress, Le Monde, radiofrance, rugbyrama, L'Equipe [2]). The second one is the transcription of the professional comments.
The texts are only in French and for one match we gathered an average of 15000 words from the press and 16 000 from the transcription.

3.2.2. Supporter corpus

The supporter written sub-corpus is also a French one for practical reasons.
It is formed of the transcription of the supporter comments, corresponding to 3 300 words for one match and a collection of blog comments about the match, about 800 words.
This corpus is characterized by colloquial terms and expression at the lexical level and syntactic disruptions at the morpho-syntactic level.

[1] http://trans.sourceforge.net/en/presentation.php
[2] 20 minutes: http://www.20minutes.fr/; Sport365: http://www.sport365.fr/; Le Figaro: http://www.lefigaro.fr/; Sport24: http://www.sport24.com/; AFP: http://www.afp.com/; L'Express: http://www.lexpress.fr/; Le Monde: http://www.lemonde.fr/; Radiofrance: http://www.radiofrance.fr/; Rugbyrama: http://www.rugbyrama.fr/; L'Equipe: http://www.lequipe.fr/ accessible on the{20080410}

The systematic comparison of these two corpora would bring us information about behaviours in sports at phonological, terminological and syntactical levels within a pragmatic approach.
Since our aim is to semi-automatically extract passages representative from emotional behaviours in sports comments, we want to use our corpus to build a typology of relevant emotional expressions in sports discourse.
To define the required parameters, we lead a two-fold-analysis: prosodic and semantico-lexical. In the next section, we present how we built a semantico-lexical model.

4. Termino-Ontological viewpoint of the sport ontology

The semantico-lexical analysis is led in taking into account several features of the term: its form, its organization, its sense(s), its use(s) – i.e. context -.
This analysis is based on now classical text analysis methods as we chose to extract terms and relation from our corpus.
Our methodology is as follow:
(1) The first step is to constitute this corpus.
(2) The second step is to extract terms from these texts and then to identify which one are the most representative of the domain (tf-idf[3] frequency and Named Entities semi-automatic tagging) and which syntactic constituents and patterns are relevant in our corpus.
(3) We are then able to extract paradigmatic as well as syntagmatic relations among terms to structure them:

a. We first identify some heuristic rules;

b. Secondly, we automatically identify relations from these rules

c. And thirdly, we structure the relevant terms

(4) We finally represent this information in formalism suitable for a Termino-Ontological Resource (TOR).
To build this TOR, we re-used and adapt some existing tools for French language parsing and semantico-syntactic corpus tagging. Extracting terms is a crucial step in the TOR building. Numerous efficient tools exist for several languages. In our French language context, tools as Termino (David and Plante 1990), FASTR (Jacquemin and Bourigault 2003) MANTEX (Frath, *et al.*, 2000), ACABIT (Daille 1999), LIKES (Rousselot, *et al.*, 1996) can be relevant for our purpose but are not strictly available or need a lot of configuration by user. Syntex (Bourigault, *et al.*, 2005) is another tool which is available and a robust analyzer. It uses linguistics resources to analyze a corpus in syntactic dependencies and gives contextual result enabling us to use patterns. Syntex permits to parse natural unstructured language as in our corpus. We then plug an extracting algorithm in its outputs. The module we developed is searching for syntactic constituents matching morpho-syntactic patterns (Lortal *et al.*, 2007). These patterns extract complex nominal phrases (nominal syntagms modified by prepositional syntagms) as well as simple verbal phrases (a verb followed by a nominal phrase). We always extract the largest covering pattern and couple this extraction with a term frequency analysis and a named entity tagging. Once extracted, as they are limited, terms are manually organized under concepts.
To build the ontological level, we based our conceptual level on an existing rugby thesaurus (Hourcade, 1998). It contains more than 6000 French and English rugby terms. We re-used about 1500 terms for structuring our TOR. But this is to be refined with a larger journalistic written corpus. In the same way, the TOR is going to widen out with a larger supporter written corpus.
This TOR is to be used as a set of tags to annotate actions during the match. In order to annotate the emotional expression in our corpus, we are manually building the emotional termino-ontological viewpoint from a fine-grained discourse analysis and the analysis of several emotion models (OCC (Ortony *et al.*, 1988), SEC (Scherer, 1988), Plutchik's model (Plutchik, 1980). We also examine the re-use of the W3C[4] emotion markup languages (Schröder *et al.*, 2007).

Serenity	hasDegree Low
	isComposedOf Optimism
	isComposedOf Love
	is Passive
Joy	hasDegree Medium
	isComposedOf Optimism
	isComposedOf Love
	is Active
Ectasy	hasDegree High
	isComposedOf Optimism
	isComposedOf Love
	is Active

Figure 1: Emotion Conceptual Level in TOR

5. Prosodic analysis and active emotion profile

For this study concerning prosodic analysis, we focused on the fundamental frequency parameters and the rhythm of speech. Intensity and energy features were rejected, because of the digitalized or noisy nature of the corpus. In the journalistic corpus provided from TV, we cannot control the recording and broadcasting conditions. Intensity and energy features may have been modified during the broadcasting processes. Moreover, concerning the supporter corpus, we also rejected intensity and energy features. The recordings were too noisy.

5.1 Fundamental Frequency

The fundamental frequency (F0) defines the pitch level of the voice. Pitch variation and contours are pertinent features for vocal studies of emotions (Bänzinger, *et al.*, 2001).

[3] Term-frequency / Inversed-Term Frequency

[4] http://www.w3.org/2005/Incubator/emotion/XGR-emotion-20070710 {20080410}

F0 measures were extracted automatically with WinPitchPro (Martin 2000). This software takes into consideration the transcriptions and signal segmentations first made with Transcriber. WinPitchPro recognises all speakers created withTranscriber and processes them separately in specific layers. F0 was extracted from all the speakers turns (at a sampling rate of 20 ms). Some statistics were performed, calculating the minimum, maximum, mean and range of F0 for each speaker turn.
The voice amplitude of each speaker, i.e. the delta difference between the maximum and the minimum of fundamental frequency, was divided in four equal registers: Low (L), Medium-Low (ML), Medium-High (MH), and High (H). The F0 values of these registers vary from speaker to another one. F0 means were calculated for each speaker turn, using F0 automatic extractions, and then each value was classified in the corresponding register.

5.2. Rhythm

We measured the speech rate, i.e. the number of syllables uttered in one second of speech (containing pauses, disfluencies etc.).
When segmenting in speaker turns and transcribing the recordings, any silent segment above or equal to 200ms is considered as silent pause within the discourse and as so, removed from the speaker turn.
The emotion felt by the viewer watching sports entertainment is not recognised as it is as an emotion. In fact, it is not one of the basic acknowledged emotion.
So, we are not able to annotate our corpus with a perceptive analysis explaining which emotion is in the speaker comments. On the other hand, following our hypothesis according to which specific game periods constitute a stimulus which leads to an emotional reaction from the speaker, then we are able to determine which prosodic parameters systematically intervene in the sports discourse for one game action.
Our hypothesis is that the emotion felt and expressed by the viewer watching sports entertainment corresponds to an active emotion profile. Concerning the prosodic parameters, an active emotion profile is characterized by an increasing of F0 values, of intensity and speech rate (Scherer, 2003). By analyzing prosodic realizations of the viewer for one game action, our aim is to verify our hypothesis and define a vocal profile of the "emotion" in sports discourse.

6. The Drop Analysis

6.1. Prosodic Analysis

In this part, we show an example of prosodic analysis made on the sportive comments realized by French journalists, during the match which opposed France against Argentina. We focused on a specific game action : the drop. The drop is "a kick made as the ball bounces after being dropped to the ground" (s.v. drop; Hornby, 1989).
We present the prosodic analyses realized on the discourses of the two main French journalists, two male speakers. Moreover, assuming that emotion expressed in the sportive discourse is a reaction to the view of the game action, we focused our analyses to the descriptive parts of the discourse. Indeed, we consider that the journalist's explanations,, concerning a definition of the drop, to the attention of the TV's viewers, are not available.
We analysed all the descriptive parts referencing to the kick-drop in the both journalists' discourses.
Each description is constituted by two to five speech segments (delimited by silent pauses), for an average duration of 6.87 seconds. So, the drop is a relatively brief game action.
Both linguistic and prosodic analyses permitted to distinguish three periods in the descriptive discourse. First, the journalist announces the game action; then he describes live the game action, this period of the description forms the main point of the discourse; finally, the effects of the game action are described and commented.

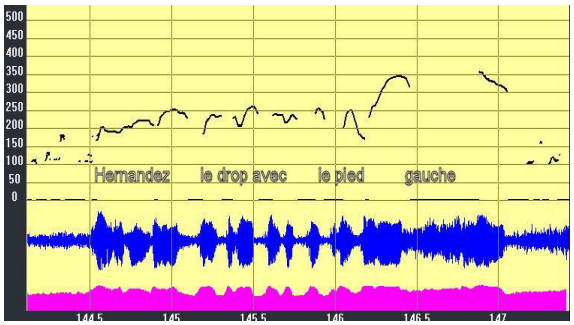

Figure 2: Pitch contour of the sentence "Hernandez le drop avec le pied gauche/*Hernandez the drop with the left foot*", corresponding to the first description period, announcement of the game action.

The figure 2 shows an example of a pitch contour for an announcement of the game action. In this first period, the action game and the name of the player, who has been designed to do the action, are announced.
We can observe that the pitch contour at the beginning of the speech segment is quite flat, contrasting with the deep rise of F0 on the penultimate syllable.
The second speech segment of this example of game action's description constitutes both the description of the drop and the results of this action.

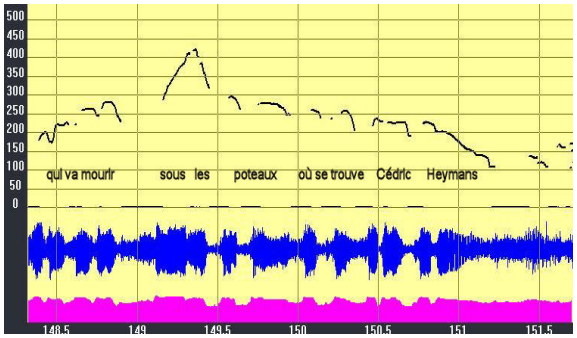

Figure 3: Pitch contour of the sentence "qui va mourir sous les poteaux où se trouve Cédric Heymans/*which goes dying under the posts where Cédric Heymans stands*"; corresponding to the second and third description periods, i.e. description of the action and its results on the game.

We observe that the F0 mean (240 Hz) and the speech rate (4.71 syll./sec.) did not vary from 1st to 2nd speech segment. On the other hand, F0 maximum increased up 425 Hz. This high pitch value corresponds to the most crucial period of the action. By contrast, the end of the sentence, describing the results of the action, shows a regular fall of F0 contour.
These examples of pitch contours show that an important prosodic variation corresponds to the period in the journalist's discourse, which describes the action being led. Furthermore, the prosodic characteristics seem to be the same that the ones relevant for active emotion profiles, as anger or joy, i.e. an increasing of F0 mean, maximum and range. Speech rate do not seem to correspond to this profile, but we can propose the hypothesis that the speech rate of the speaker, in the description periods of discourse, follows the conduct of the game action, rather than expresses the speaker's emotion.

6.2. Semantico-lexical analysis

We have two main perspectives on our semantico-lexical analysis. The first one is a terminological analysis and the second an emotional analysis (emotion marks in discourse).

6.2.1 Terminological analysis

The two corpora have a lot of meta-discursive discourse during the comments. The speakers explain the terms they use to comment, and even comment them. The analysis permits us to retrieve the terms, mostly expressions, which slip through the net of our patterns. In (1) the speaker uses a simple pattern for picking out synonyms, using "qu'on appelle aussi / *that we also call*" to link "coup de pied tombé/literally *felt kick-drop kick*" with "drop/*drop*".

(1) speaker#1: coup de pied tombé pour ce renvoi qu'on appelle aussi drop
speaker#1: drop kick for this drop out that we also call drop

To retrieve new terms, "rugbalistic" terms as French rugbymen say, the manual analysis is compulsory as it often come with long comments. For example the hierarchy (2') is coming fro the analysis of the sequence (2).

(2) speaker#2: hé bé vous jouez au pied
speaker#2: hey bey you play with foot[5]
speaker#2: ou des chandelles c'est-à-dire en l'air la quille
speaker#2: or up and under kick that is to say the skittle
speaker#1: c'est la chandelle c'est votre jargon rugby
speaker#1: up and under kick that is your rugby jargon
speaker#1: d'ailleurs c'est très bon hein
speaker#1: besides it is really good (hein)
speaker#2: la quille c'est la chandelle ou alors un drop ce que vous avez vu par Contepomi
speaker#2: the skittle is the up and under kick or also a drop what we saw with Contepomi

(2')

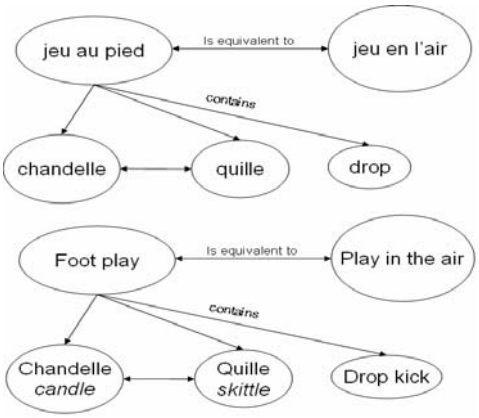

Figure 4: Rugby TOR (part)

In future, we hope, thanks to terminological analysis, to compare the terms between specialist in professional context and supporters in leisure context. The second perspective on this analysis is to be able to recollect corpus year after year, since each seasons is marked by the arrival of new terms as *ovalie* and *terre d'ovalie* – meaning something as *ovaly country* and *land of oval sport*- some years ago. Today these terms are accepted and used for marketing and as a community term (among rugbymen).

6.2.1 Emotion mark analysis

The lexical analysis of emotion is really different from the written journalistic corpus and the oral or written supporter corpus.
As the first one is descriptive and public, we can find terms about what the speaker feels or what the players feel. We can find terms as: tomber de haut/ *to fall headlong*, désillusion/ *disillusion*, doucher les ambitions/ *to tell off ambitious*, trop nerveux/*too nervous*, tendu/*tense*, tétanisé/*tetanized*, intentions offensives/*offensive intentions*, jeter un froid/*to cast a chill*, perturbé/*upset*, faible/*weak*, volontaire et enragé/*headstrong and keen*,…
Another analysis on emotion to be done is the morpho-syntactic analysis (Kerbrat-Orecchioni, 2000). We also observed that a typographic analysis may also leads us to useful patterns of emotion expression as shows (9). The typography used is representative a means of emotional expression in written comment.

[5] nota: you do not run the ball

(9) ROUGERIIIIIIIIIE !!!
ROUGERIE !!!
Willx est très en colère. Les 17 points encaissés en une mi-temps, il a pas aimé!
Willx is really angry. The 17 hit points in one half-time, he *don't* like it !
VOILA!!
HERE WE ARE!!
PUNITION- Corleto !!!!
PUNISHMENT – Corleto !!!
The second one is really much more rude and is about what the speaker feels. Examples from (1) to (8) show the terms we can find.
(1) speaker#1: ah c'est pas bien ça les Français de siffler comme ça mais c'est logique
speaker#1: ah it's naughty that he French to whistle this way but it's logical
(2) speaker#1: ben hé qu'est-ce qu'il fait non il dit rien là ben si ah ben quand même
speaker#1: ben hey what is he doing no he is saying nothing ben yes ah ben well really
(3) speaker#1: putain l'autre lui passe par dessus alors quelquefois je peux être ordurier quand même hein
speaker#1: fuck the other is passing above so sometime I can be filthy well really hu
(4) speaker#1: ouais mais va falloir taper là les mecs putain les mecs il va falloir quand même qu'ils se décident à aller
speaker#1: yeah but should hit now guys fuck guys they should decide on to go finally
(5) speaker#1: pas vrai ça
speaker#1: not possible
(6) speaker#1: qu'est-ce qu'ils font là les mecs ils
speaker#1: what are they doing now guys they
(7) speaker#1: qu'est-ce qu'il y a comme foot hein
speaker#1: so much football hu
(8) speaker#1: euh il a pas mis le pied en touche le mec là
speaker#1: hey he didn't put his foot in touch the guy here

The terms we find are not crucial to add to a terminology and extend the coverage our TOR. However, these speakers turns, when combine with a prosodic analysis strengthen the hypothesis of (Mathon, 2007) saying that when the lexicon is strongly marking an emotion, the prosodic parameters are lessened. The validation of this hypothesis underlines the asset of an automatic extraction of sequences.

7. A tool for ontology based indexation of segments of signal

The prosodic analysis led on a collection of manually extracted segments of sport action and providing us with an active emotion profile is to be refined by further analysis on a larger automatically extracted corpus. However, we can already explain how we will use the built TOR to index segments of signal by means of a corpus analysis tool enabling lexical, prosodic and discourse analysis.
The first requirement for such a tool is to have prosodic processing functionalities. That's why we base our module on Winpitch (Martin, 2000). WinpitchPro allows real time monitoring of recordings (spectrogram, Fo, etc.), high precision segmentation, speech turns overlapping, assisted alignment of existing transcription, automatic building of speech segments database (XML output), prosodic morphing. WinpitchPro can also process multimedia files (audio/video or text with its corresponding signal).

Figure 5: Video/Prosody/Signal/Transcriptions alignment

We aim at optimizing WinpitchPro with functionalities as:

- Visualization of already aligned corpus (vidéo/signal/texts) (Fig.5)
- Speaker turns synchronisation based on the signal time
- Automatic and on the fly prosodic parameters calculus based on syntactic groups (not turns)
- Segment extraction based on defined on multi-level profiles (lexical - semantic - prosody levels)

At the moment, we are refining these specifications and launching a global project on multi-linguistic level analysis tool.

8. Conclusion

We presented here a preliminary study aiming at defining lexically and prosodically the emotion or the emotions provoked by a sport show. These emotions are expressed in both journalistic and supporter discourse
The prosodic level analysis enables us to say that the emotion expressed in speech sequences describing a specific action in the game has parameters similar to an active emotion. This observation leads us to envisage the detected prosodic parameters as evidences of the speaker excitement when the action occurs.
While we do not deem our results as definitive or universal, we wish to use them as prosodic parameters associated to lexical parameters in corpus processing. The association of these prosodic and lexical parameters with emotion concepts considered

as universal creates a large and fine-grained termino-ontological resource. The TOR is used to support multimedia documents annotation. The fine and automatic annotation of an oral corpus will enable us going further in the automatic extraction of representative passages. It represents an important saving of time and a finer linguistic analysis of our journalistic sports discourse corpora and its expressivity marks.

We still have a lot of question about the variation of emotion provoked by the sports show. Does it vary with the game actions, with the spirit of the spectator (pro or cons the winning team), or with the expressed spectator mood (terms in use)? Is this emotion in sports is universally expressed or culturally and linguistically dependant? Our corpus is still shallow-analysed and we hope to dig it thanks to a cross domain analysis.

9. Acknowledgements

We want to thank Ph. Martin, Professor at the Paris VII University for WinPitchPro, the Rugby Club of Versailles for providing us with supporters, and the friends all around the world for video tape recording.

10. References

Bänzinger, T., Grandjean, D., Bernard, P. J., Klasmeyer, G. & Scherer, K. R. (2001), Prosodie de l'émotion: Étude de l'encodage et du décodage, in *Cahiers de linguistique française* 23, pp. 11-37.

Bourigault D., Fabre C., Frérot C., Jacques M.-P. & Ozdowska S. (2005), Syntex, analyseur syntaxique de corpus, in *Actes TALN 2005*, Dourdan, France

Campbell, N., (2001), The recording of emotional speech (JST/CREST database research), in *Proceedings from COCOSDA Workshop 2001*, Taejon, Korea.

Chung, S.J., (2000), L'expression et la perception de l'émotion extraite de la parole spontanée : évidences du coréen et de l'anglais, Thèse de l'Université Paris 3, 264 pages.

Daille B., (1999), Identification des adjectifs relationnels en corpus in *Actes de TALN,* Cargèse.

David S. et Plante P. (1990), De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. In *Intelligence Artificielle et Sciences Cognitives au Québec*, 3(3):140-154

Douglas-Cowie, E., Cowie, R., Schröder, M., (2001), A new emotion database: considerations, sources and scope, in *Proceedings from ISCA Workshop 2000*, Newcastle, North Ireland.

Fónagy, I., (1983), *La vive voix: Essais de psycho-phonétique*, Bibliothèque scientifique, Payot, Paris.

Frath P., Oueslati R. and Rousselot F., (2000), Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques, in *Ingénierie des connaissances. Évolutions récentes et nouveaux défis*. J.Charlet, et al. (eds). Eyrolles, Paris, pp 291-304

Hornby, A.S., *Oxford Advanced Learner's Dictionary of Current English Oxford*, England: Oxford University Press, 4th edition, 1989

Hourcade, B., (1998), *Dictionnaire Du Rugby,* La Maison du Dictionnaire (ed), 220 pages.

Jacquemin C. and Bourigault D. (2003), Term Extraction and Automatic Indexing, in Mitkov R. (ed), *The Oxford Hand-book of Computational Linguistics,* Oxford Univ. Press, pp. 599-615

Kerbrat-Orecchioni, C. (2000), "Quelle place pour les émotions dans la linguistique du XXe siècle? Remarques et aperçus", in *Les Emotions dans les interactions*, Ch. Plantin, M. Doury, et V. Traverso (éd.), Presses universitaires de Lyon, p. 33-74.

Kerbrat-Orecchioni, C., (1990), *Les interactions verbales*. Armand-Colin, Paris.

Léon. P. R., (1993), *Précis de Phonostylistique: Parole et expressivité*. Nathan, Paris.

Lortal G., Todirascu-Courtier A., Lewkowicz M., (2007,) AnT&CoW: Share, Classify and Elaborate Documents by means of Annotation, in R. Chbeir, P. Pichappan, A. Abraham (eds), *Journal of Digital Information Management*, pp. 61-70.

Martin, P., (2000), WinPitch 2000: a tool for experimental phonology and intonation research, in *Proceedings of the Prosody 2000 Workshop*, Kraków, Pologne, 2-5 October 2000 http://www.winpitch.com/ {20080410}

Mathon, C., de Abreu, S., (2007), Emotion, from speakers to listeners. Perception and Prosodic Characterization of affective speech, C. Muller, S. Schotz (eds), *Speaker Classification II, Selected Projects*. LNCS, Springer, pp. 70-82.

Mathon, C., (2007), Multimodal Analysis of Anger in Natural Speech Data, in J. Trouvain and W. J. Barry (eds), in *Proc. of XVIth International Conference of Phonetic Sciences*, pp. 2117-2120.

Plutchik, R., (1980), A general psychoevolutionary theory of emotion. In Plutchik et Kellerman (Eds.), *Emotion: Theory, research, and experience*: Vol. 1. Theories of emotion. pp. 3-33

Rousselot, F., Frath, P., and Oueslati, R. (1996), Extracting concepts and relations from Corpora. In *Proceedings of the Workshop on Corpus-oriented Semantic Analysis*, European Conference on Artificial Intelligence, ECAI 96, Budapest.

Scherer, K. R. (2003), Vocal communication of emotion: A review of research paradigms, in *Speech Communication*, 40, 227-256.

Scherer, K. R., (1995), How emotion is expressed in speech and singing, in *Proceedings of the 13th International Congress of Phonetic Sciences.* Stockholm, Sweden, pp. 13-19.

Scherer, K.R. (2000), Psychological models of emotion, in J. Borod (Ed.). The neuropsychology of emotion, Oxford/New-York: Oxford University Press, pp. 137-162.

Schröder, M., Zovato, E., Pirker, H.,Peter, C., Burkhardt, F., (2007), W3C Incubator Group Report.

Flames, Risky Discussions, No Flames Recognition in Forums

Maria Teresa Pazienza[a], Armando Stellato[a], Alexandra Tudorache[ab]

a) AI Research Group, Dept. of Computer Science, Systems and Production
University of Rome, Tor Vergata
Via del Politecnico 1, 00133 Rome, Italy
{pazienza,stellato,tudorache}@info.uniroma2.it

b) Dept of Cybernetics, Statistics and Economic Informatics, Academy of Economic Studies Bucharest
Calea Dorobanților 15-17, 010552,
Bucharest, Romania
alexandra.tudorache@gmail.com

Abstract

In this paper we describe our experimental study on flames and risky topic recognition. Firstly we introduce our approach, experiments and research goals. Then we provide basic definitions of flame, risky discussions and their difference. Furthermore we will analyze the most important features of flames by highlighting general, Italian and English features. Then we will focus on experiment and corpus description concluding by commenting the results of our test.

1. Introduction

World Wide Web supports the creation of virtual communities and interactive websites at a large extent. Forums and discussion boards represent a very important and significant social phenomenon inside the WWW. The traditional face to face discussions are migrating into the Web with new and limited rules for maintaining person to person interaction in a polite mainstream. Previously unknown persons discuss about several topics, while each participant contributes to the discussion with his own background, culture and language comprehension: totally uncertain contexts emerge!

Moreover, in those environments, traditional important elements of communication (such as gestures and voice tone) that underline emotion and sentiments are missing. Sometimes they are replaced by *emoticons* (icons that express emotion) but, for example, they do not wear voice inflections. As a consequence misunderstandings and flames occur in forums and discussion boards at an higher rate than in face to face conversations. Moderators are forced to block discussions between involved persons and interaction fails.

To maintain democracy on the web while avoiding unpolite and violent discussions, it is important for forum and discussion groups administrators to be able to recognize possible source of flames before they reach an undesirable state, without the need of constantly checking whole message boards.

We have analysed several forums in order to verify the existence of specific contexts enabling us to recognize the phenomenon. In the literature, at a border line between linguistics, psychology and cognitive science, we found a few contributions that have been revisited and enriched with classification algorithms in a computational approach.

Multidisciplinary studies have been considered to develop our approach and set up an experiment.

As we will use a specific terminology for which there not exists still a common understanding, by first we will provide definitions for specific terms object of our analysis. Then the adopted methodology, the test corpus and results will be widely discussed.

2. Flame, Forum, Risky Discussion, Topic and Post Definitions

A *forum* is a virtual "place" that hosts several discussions considered as written conversations. Forums can be either generic (where discussions span over many topics) or specific boards (where participants can talk about only one domain like politics, art, religion etc).

We define the *topic* as one discussion thread with one subject and more interactions (posts).

A *post* is the response provided by a specific user. Usually forums require login and it is not possible to post anonymously.

A *flame* is defined by a sequence of "non constructive" posts, with no positive contribution to the discussion. In flames users attack each other at a personal level instead of contrasting the dialog partner for his/her approach, contribution to the discussion or argumentation. As a consequence the flames phrasal structure is closer to oral than to written dialogue. Flames often induce moderators to close discussions.

We do a neat distinction between flames and risky discussions; in fact *risky discussions* contain a few flaming elements but could not be categorized as flames. They express a sort of "expectation" for a flame!

In real life the distinction between flames and risky discussions is usually subjective. It depends on moderator skills, his attitude, level of stress and level of implication in the subject. Sometimes the moderator himself generates flames due to his tough policy.

For our experiments we recognize a flame as a sequence of

six to eight posts that disrupt topic and represent personal attacks, while a risky discussion is considered as a sequence of three to four posts of personal attacks.

3. Flame and Risky Topics Features

We widely analyzed several forums either in English or in Italian languages in order to identify the existence of "flame features". By accessing also to a few psychological studies (King, A., 1995 ; Bucci, W., Maskit, B., 2005; Leahy, S., 2006) we have been able to highlight specific linguistic behaviors in flames. To start we collected language independent features, then language specific ones (Italian and English). It seems, in fact, that differences in cultures determine different ways for flaming.

General flame features:

1. Discussions take place between two or maximum three users on a specific topic or over more topics.
2. Short posts without much argumentation.
3. Posts made by new users (newbies) that are not able to integrate from the beginning into that specific community. (King, A., 1995)
4. Offensive language.
5. Phrase or expression ambiguity (eg. "un" against "il"; "bene", "niente" – Italian, "the" against "this", "the idea" against "my idea"- English); psychological clinical studies demonstrated that an ambiguous language is sign of tension. (Bucci, W. , Maskit, B., 2005)
6. Tough or nonlinear moderation policy.
7. Nonsense and off topic posts written by moderators in a no flame or risky discussion.
8. Off topic personal questions about other users of that forum; simple off topics usually doesn't determine a flame.
9. Non acceptance of community rules; it is related to few users that some times disrupt the activity of an entire forum. (Leahy, S., 2006)
10. Speed in posting; flames often take place almost in real time.
11. Repetitive cites from other posts. One user try to attack another user attacking every idea of that specific user.
12. Proper names are not relevant in a flame.

Italian flame features:

1. Users addresses directly to each other ("cut and thrust" - "botta e risposta" in Italian).
2. A few users attack each other frequently over more topics.
3. Use of short phrases, often without a subject (è una cosa stupida; non è vero - "it is a stupid idea; it is not true")
4. Hypocritical politeness in expressions like: "perdonami se...", "scusa se..." ("excuse me...")

English flame features:

1. Users attacking ideas instead of other users directly. (This is often not valid for sport/games and teens forums – see Notes below)
2. Users adopt a lot of ironic expressions instead of direct expressions.
3. Adoption of slang or urban expressions.

Notes: Language in sport/games and teens forums is usually biased. In these contexts, flames are recognized by user behaviors more than by language models.

Risky topics features:

We are interested also in recognizing possible risky situations, then we tried to find general contexts in which they occur. In the experimental set up, they are used as a third class in between flames and no-flames; in fact risky situations, while not being flames, reveal a few features overlapping with those of flames.

1. As in flames, arguments take place between two or maximum three users on one topic or more.
2. Risky discussions are characterized by mini-flames (two or three flame posts) followed by normal discussion.
3. In risky discussions we can find more than one mini-flame; moderators usually don't make hush interventions.
4. In risky discussions are present impersonal constructions against personal construction as in flames; when a risky topic turns personal flaming occurs.
5. In risky discussions often we will find a good number of off-topic posts.

4. Algorithms for Flames and Risky Topics Classification

In the previous section we introduced flame and risky topics features. In this section we will see some methods for identifying flames and risky discussions in forums. There are two main techniques for emotion and sentiment classification, roughly: symbolic and machine learning techniques. The symbolic approach may use manually defined rules, lexicons, conceptual knowledge,… where a machine learning approach uses unsupervised, weakly supervised or fully supervised learning to construct a model from a training corpus. (Basili, R., Moschitti, A., 2005)

Once agreed on flame feature classes (as naively summarized into previous sections) and completely structured them, we will set up a dedicated knowledge based architecture to recognize flames in forums. In fact, while we analyze discussions in forums as a textual sequences, posts contain a lot of typical oral/gergal expressions. There is a mixture of written oral structures and components thus requiring a dedicated processing.

Meanwhile we are verifying the possibility of using

machine learning algorithms as a first approach.
The main applications of supervised machine learning are classification algorithms. We will focus on document classification.
For document classification we can use classic supervised learning techniques (e.g. Support Vector Machines, Naive Bayes, Maximum Entropy).
An efficient yet simple classification algorithm is Naive Bayes; it does the naive assumption of document features independence. Later on this document we will present some experiments on two and three document category classifier.
Unlike in supervised learning, in case of unsupervised learning methods is not needed the manual labeling of inputs (the so called training set).
A well known unsupervised method is clustering that not always is probabilistic.
A method that allows to vary the number of clusters with the problem size and also gives to the user the control over each cluster member similarity degree is the adaptive resonance theory (ART).
ART networks are used for many pattern recognition tasks. The first version of ART, "ART1", was developed by Carpenter and Grossberg in 1988.(Manning, C. D.; Raghavan, P.; Schütze, H., 2008)

4.1. Our Approach

For our experiments we used the Bayesian Classifier with Laplace smoothing supervised approach.
Naïve Bayes algorithm is based on the assumption of events independence. Although it is an "incorrect" assumption that documents features are completely independent, the Naive Bayes algorithm proved over time to have a very good performance in text classification (Basili, R., Moschitti, A., 2005; Manning, C. D.; Raghavan, P.; Schütze, H., 2008; Yu Bei; Unsworth J., 2007). Moreover, many of our analyzed features (e.g. length of the post, newbie recognition etc...) relate to information which goes out of the textual boundaries of the documents and is thus more prone in being used with such assumption.

4.2. Experimental Setup

For our experiments we adopted the "Waikito Environment for Knowledge Analysis" Weka set (Witten, I. H.; Frank, E., 2005) for using a Naïve Bayes algorithm implementation and an extension of Word Vector Tool for producing word vectors by extracting data from Forum sources.
By default Naïve Bayes in Weka's implementation uses Laplace smoothing: always adds 1 to the number of different values for a particular attribute.

5. Experiment Description

In this experiment we will verify the possibility to classify flames *versus* noflames as well as risky discussions. In fact, forum administrators are interested to "prevent" (moderate risky discussions) rather than to "cure" (close flame topics).

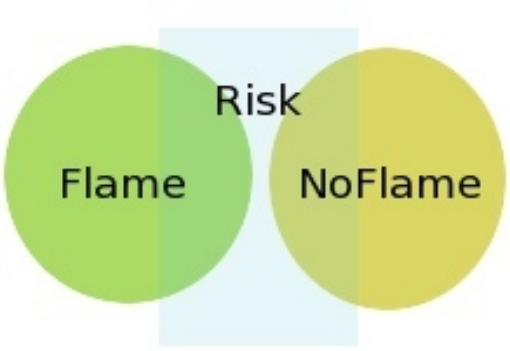

Figure 1: Topic Class Diagram

We designed two sets of experiments: one to recognize flames and another to recognize either flames or risky discussions.
Flames and noflames are two disjoint classes while (as discussed in section 3) risky topics class has features from both flame and noflame classes thus representing a sort of fuzzy class.
For training purposes risky topics were extracted from noflame category meanwhile flame category remains unchanged during both tests.

5.1. Corpora Structure

From observing several forums and discussion boards we learned that flames usually are generated by few users that express their ideas using a rather aggressive language. Moreover each forum is a community with its own rules, participants, language, interests, moderation and administration staff. Each of these elements impact significantly on flames characteristics.
For data homogeneity and evaluation purposes, we must use the same forum both for training and testing. The Italian corpus is represented by sequences of posts extracted by topics found on ***postare.it***. We choose this forum as it is general, rich of politics and daily matters and has the right content structure for our experiments: normal/noflame discussions (65%), flames (5%) and risky discussions (30%).
In forums (as in any virtual or real community) the level of emotionality changes over time and so does the forum structure. Hereafter we provide a few examples from training corpus in Italian language.
All the examples have been translated into English preserving as much as possible of the original sense and style to better illustrate emotion. We tried to maintain phrase structure and grammatical non concordances, too.
" by Gregoriosurace
Testo Quotato:Postato originariamente da Ilpiero :
Sfogo amaro condivisibile ma...
Testo Quotato:da GregorioSurace: Complimenti comunisti.

- mi spieghi cosa c'entra? Senza polemiche o altro... A una persona che mi chiede una delucidazione non posso fare altro che dargliela. Evidentemente il mio testo non era abbastanza chiaro. Anche se a me sembrava cosi...” -
“I understand your bitterness but...
Quoting Text: GregorioSurace: Compliments to Communists. - Could you explain? No polemics please... If someone asks me something, I must answer. Obviously my post wasn't clear enough. Although it seemed to me so ... ”
(Excerpt from training **flame** 1779.txt)
“by Francesca
---
tutti noi abbiamo la nostra libertà di pensare ciò che vogliamo di chiunque.
ma c'e anche da dire che io non ti volevo convicere in nulla io ti stavo solamente dicendo che come tu non hai bisogno di niente, meglio per te tu hai il tuo stile di vita ma ognuno di noi hanno dei vizi chi il fumo, chi alcol, chi spendere i soldi, chi riempire il cu+o allo stato combattendo le loro battaglie e poi questo e il mondo in qui viviamo, un mondo di guerre combattute dai poveri e in tanto lo stato si fa i villoni e i viaggietti con l'aereo privato...ma lasciamo perdere....
la droga esiste da sempre...secondo te gli indiani nella pipa della pace cose c’era tabacco? anche nel fumo ci sono delle regole da rispettare e se propio lo voi sapere da come gira il mondo ora preferisco uno che si fa una canna al giorno che uno che crede un qualcosa che non e!

_____
by Asmodeus

E tientelo pure guarda, fai un favore a tante altre persone che invece non lo vogliono “
“By Francesca
---
*We all are free to think what we want of anybody. But I also want to say that I didn't want to persuade you on anything. I was just saying that much the same way you don't need anything, it is much better for you, you have your own lifestyle but everyone have their vices like smoking, drinking, squandering money, fill the state *** fighting their battles and thus this is the world we live in, a world of wars fought by the poors and where the state builds villas and travels by private jets... but let's go over it ....*
Drugs have always existed ...do you believe in the Indian peace pipe there was tobacco? There are rules to be respected even for smoking and if you'd like to know my opinion I prefer a person who smokes drugs once a day to one who believes in something that is not!

_____
By Asmodeus----Look do us a favor and keep it”
(Excerpt from training **risky** discussion r_2151.txt)
As it is evident, there is a very narrow difference between flames and risky topics and this is reflected in the experiment results as we will see in results section.

5.2.Training and Testing Notes

The annotation of each sequence of topics was made manually. As stated in section 3 flames were extracted from integral topics as 6-8 flame post sequences, risky discussions as sequences of 6-8 post in which 2-3 post sequences were mini-flames.
Testing was made on the same forum. Corpus was preprocessed as for the training. After manual classification the training and testing documents were selected randomly to guarantee a fair test.

5.3. Results and Evaluation

For evaluation purposes, as in information retrieval, the two main indicators - precision and recall - have been used.

$$Recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

$$Precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

For our experiments we used a weighted precision and recall.
In forum administration is preferred to recognize all flames and have some false positives than to not completely recognize flames.
We considered “error weight” as an index of the gravity of the error. Flames recognized as noflames and noflames as flames errors are considered “full errors” and have error weight=1. Risky topics recognized as flames and risky topics recognized as noflames are considered “partial errors” and we attributed them an error weight of 0.5.
So for each experiment precision and recall formulas have been rearranged as in the following:

$$Recall = \frac{tp}{tp + \sum fn \times ew}$$

$$Precision = \frac{tp}{tp + \sum fp \times ew}$$

Where:
tp = true positives - topics correctly identified
fn = false negatives - correct topics that have not been found
fp = false positives – incorrect topics classifies as positives
ew = error weight

6. Experiment Results

In this section we analyze results of a couple of experiments in flame, no flame and risky topics identification.

6.1. First Experiment Results: Flames and No Flames Classifier

In the first experiment we were interested in flame and no flames classification.
The training set was composed by 35 flames and 95 noflames. The testing corpora was composed by 12 flames and 97 noflames including risky topics. Test results are shown in Table 1.

Flames identified as noflames	0
No flames identified as flames	18
Risky topics identified as flames	3
Flames correctly identified	12
No flames correctly identified	76
TOTAL FLAMES TESTED	12
TOTAL NO FLAMES TESTED	97
TOTAL TOPICS TESTED	109

Table 1: First experiment results: flames and no flames recognition

As you can see in Table 1 we identified all flames and a good number of noflames - 76 from a total of 97: about 78% .
In Figure 2 we presented the correctly identified topics by category.

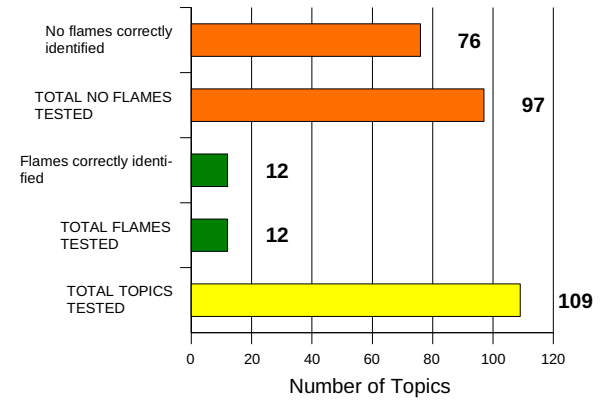

Figure 2: First Test Correctly Identified Flames

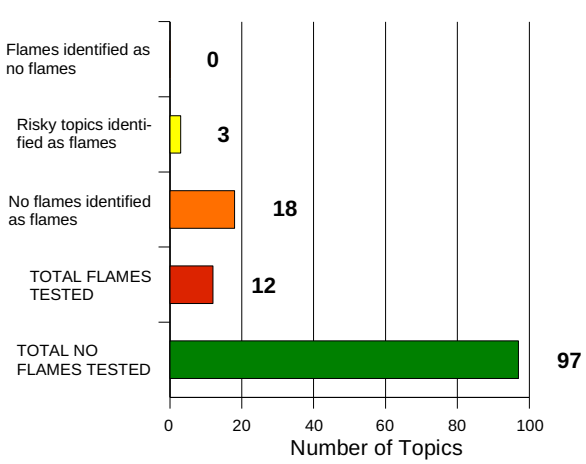

Figure 3: Errors in Topic Classification

In Figure 3 you can see the errors of topic classification. In conclusions and future work section we will present some methods to minimize such errors.
While we obtained a complete recognition of flames, we get some false positive b:most of them could be classified as risky discussions by a manual inspection.
E.g.
" by Mime
....appena sentito al tg3.... il nanetto "non c'e stato alcun editto bulgaro, il mio era un appello ai nuovi dirigenti che stavano per insediarsi in rai che "certe cose" non si verificassero piu"
mi domando e dico esistesse un limite all' idiozia e alla sfrontatezza di quell' omuncolo?
criminoscio "
"By Mime
.... Just heard on tg3 .... The little one "there 'was no Bulgarian edict, mine was just an appeal to the new leaders coming into RAI... that" certain things "shouldn't happen any more"
I wonder if there is a limit to stupidity and impudence of that little man! Criminoscio" (Excerpt from ***risky*** *discussion r_1705.txt)*
The two main indicators are the following:

Recall	1
Precision	0.53

Table 2: First Experiment – Flame Recognition Precision and Recall

Precision and recall indicates that on this first classification we have the ability to find all flame topics and that the probability to find the most relevant topics first is over 50%. From a forum administrator it could be a good result: in fact forums are very dynamic contexts and used language is not

standard at all - each participant uses his own language, abbreviation ways to express disappointment etc.
Other studies have reached a lower recognition rate like Ellen's Spertus Smokey. (We have not been able to find further references on flame recognition). In Table 3 we reported the comparative results of our test based on Naïve Bayes Smoothed and her tests based on C 4.5 algorithm. (Spertus, E. , 1997)

	Naïve Bayes Smoothed	**C 4.5**
Flames Recognition (%)	100%	39%
No Flames Recognition (%)	78%	97%

Table 3: Comparative Results of Naive Bayes and C 4.5 Algorithms

6.2. Second Experiment Description: Flames, Risky Topics and No Flames Classifier

For the second experiment we annotated three classes: flames, noflames and risky discussions.
The training set was composed by 35 flames, 95 noflames and 22 risky discussions. The testing corpus was composed by 12 flames, 88 no flames and 9 risky topics.
Once again we get no false negatives on flames/noflames and only two false negatives on flames/risky topics identification as shown in Table 4.

FLAMES	
Flames identified as noflames	0
Flames identified as risky topics	2
Flames correctly identified	**10**
No flames identified as flames	19
Risky topics identified as flames	3
No flames+risky topics correctly identified	***75***
TOTAL FLAMES TESTED	**12**

Table 4: Flame Topics Identification

After a manual inspection, the two flames' false negatives classified as risky discussions revealed to be not very aggressive ones. It is remarkable that no flame was identified as noflame.
The risky topics identification is the weakest test: it is not a surprise due to the nature of risky topics. Risky topics have both elements of flames and of noflames topics and by definition it is not a disjoint class as shown in Figure 1.
For risky discussions identification we have only one true positive from 9 tested, 3 topics were identified as flames and 5 as noflames (see Table 5).

RISKY TOPICS	
Risky topics identified as flames	3
Risky topics identified as no flames	5
Risky topics correctly identified	**1**
Flames identified as risky topics	2
No flames identified as risky topics	2
No flames+flames correctly identified	***77***
TOTAL RISKY TOPICS TESTED	**9**

Table 5: Risky Topics Identification

No flames are correctly classified in a proportion of 76%. On noflames we have 21 false negatives of which 2 topics are identified as risky discussions.

NO FLAMES	
No flames identified as flames	19
No flames identified as risky topics	2
No flames correctly identified	**67**
Flames identified as noflames	0
Risky topics identified as noflames	5
Flames+risky topics correctly identified	***11***
TOTAL NO FLAMES TESTED	**88**

Table 6: No Flame Topics Identification

We had 5 false positives but only risky topics identified as noflames.
In Tables 7,8 and 9 we show precision and recall for every class.

Flames Recall	91%
Fames Precision	33%

Table 7: Flames Classification Indicators

Risky Topic Recall	20%
Risky Topic Precision	33%

Table 8: Risky Topics Classification Indicators

No Flames Recall	77%
No Flames Precision	96%

Table 9: No Flames Classification Indicators

Flames Identification Recall raises to 91%. Risky topic identification instead has a poor precision and recall as expected.

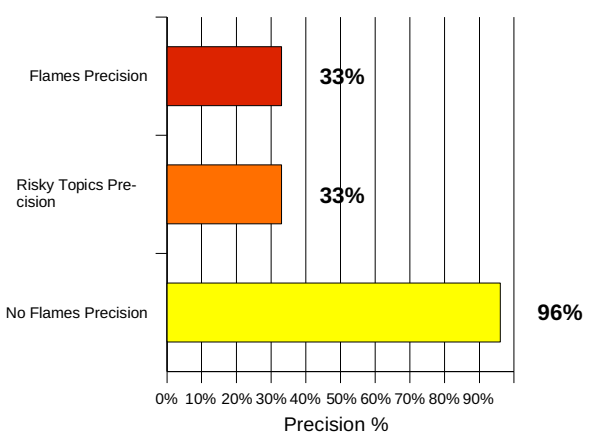

Figure 4: Second Test Comparative Precision

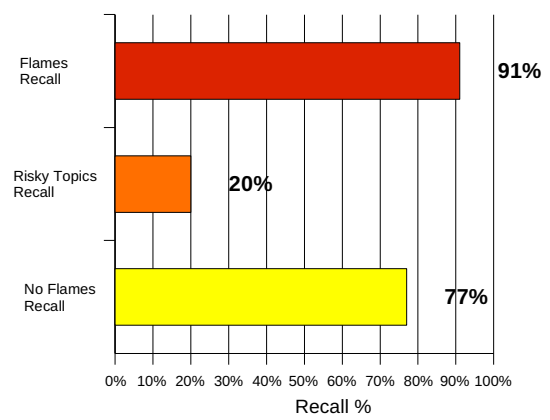

Figure 5: Second Test Comparative Recall

In Figures 4 and 5 we presented the comparative results of precision and recall for the second experiment.
Results clearly show the difficulty of risky discussions categorization and the duality of risky topics. Each topic has both elements of normal – no flame and flame topics.
Since risky topics present heavily dependent features their identification goes against Bayes' assumption of independence. So we need other methods to optimize results. A possible approach will be presented in Conclusion and future Work section.

7. Conclusions and Future Work

In this paper we presented our experimental study on flames and risky topic recognition later on definitions of flame, risky discussions and differences between flames and risky discussions.
Then we provided an exemplification of the most important features of flames and risky discussions distinguishing between general, Italian and English ones.
As widely described, we have got very good results in flames recognition and promising ones concerning risky discussions identification. At the moment we are at an early stage in our research: we aim in searching further algorithms in order to better recognize risky situations. In fact forum administrators could benefit a lot by an early alert! The matter is very hard to formalize! We analyze discussions in forums as a textual sequences, while posts contain a lot of typical oral/gergal expressions. There is a mixture of written oral structures and components requiring a dedicated processing.
From our empirical observation we can verify that the easiest way to eliminate false positives is to identify impersonal and personal phrase structure.
Impersonal constructions identify risky discussions whereas personal phrase constructions identify flames. Impersonal construction like “it is know”, “is said”, “si dice”, “non e detto” indicates tension but not a direct attack to another user.
*E.g. “tutti noi abbiamo la nostra libertà di pensare ciò che vogliamo di chiunque.” - “we all are free to think what we want of whom we want” (Excerpt from **risky** discussion n. 2151)*
*“non per stare sempre a parlare di lui ma il gesto del lancio di uova sembra l'inizio di una protesta che covava da tempo o un gesto isolato” - “I don’t like to talk only about him, but the act of trowing eggs seems the beginning of a long underlaying protest or an isolated act”.(Excerpt from **risky** discussion n. 1803).*
It emerges that in both sentence fragments predominate impersonal constructions while tension is sensible. These two are considered risky topics because they can degenerate easily in flames. Flames are characterized by personal constructions.
E.g. *“ma tu, hai una connessione wireless e per portare il tuo livello produttivo ai valori che cosi bene conosciamo...” - “but you, have a wireless connection and to bring your productivity to the levels we all know... “ (Excerpt from **flame** discussion n. 1651).*
“ah intendi dire che con quel post ... disturbavo qualcuno??”-”you want to say that with that post I was annoying someone?” (Excerpt from flame discussion n. 1826).
The heuristics about phrase constructions could be effective in distinguishing between flames and risky discussions.
In the future we plan study the possibility to integrate such heuristics in topic classification.

8. References

King, A. (1995) *Effects of Mood States on Social Judgments in Cyberspace: Self Focused Sad People as the Source of Flame Wars*, Storm, July 2, 1995, http://psychcentral.com/storm1.htm

Bucci, W.; Maskit, B. (2005, 1). *A weighted dictionary for Referential Activity. Computing Attitude and Affect in Text;*, 49-60

Leahy, S. (2006) *The Secret Cause of Flame Wars*, Feb 13 , http://www.wired.com/science/discoveries/news/2006/02/70179

Basili, R.; Moschitti, A. (2005) *Automatic Text Categorization: From Information Retrieval to Support Vector Learning*, Aracne Editrice, Informatica, ISBN: 88-548-0292-1

Manning, C. D.; Raghavan, P.; Schütze, H. (2008) *Introduction to Information Retrieval*, Cambridge University Press (to appear in 2008), http://informationretrieval.org

Boiy, E.; Hens, P.; Deschacht, K.; Moens, M. F. (2007) *Automatic Sentiment Analysis in On-line Text*, Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria – June 2007

Dave, K.; Lawrence, S.; Pennock, D. M. (2003) *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.* In Proceedings of WWW-03, 12th International Conference on the World Wide Web, ACM Press, Budapest, HU, 2003, pp. 519–528.

Yu Bei; Unsworth J. (2007) *An Evaluation of Text Classification Methods for Literary Study*, University of Illinois at Urbana-Champaign, http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=157

Janssen, Jerom F., (2007) *Diachronical Text Classification, A study of text properties and their changes over time*, PhD. Thesis, University of Groningen, May 14.

Weka, Weka 3: *Data Mining Software in Java*, http://www.cs.waikato.ac.nz/ml/weka/

Witten, I. H.; Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (Second Edition). Morgan Kaufmann

Word Vector Tools, *The Word & Web Vector Tool*, http://nemoz.org/joomla/content/view/43/83/lang,en/

Spertus, E. (1997) *Smokey: Automatic recognition of hostile messages.* In Proceedings of the Ninth Conference on Innovative Application o f Artificial Intelligence (IAAI-97), Providence, RI, pages 1058-1065. AAAI Press/The MIT Press, July 1997

Automating opinion analysis in film reviews : the case of statistic versus linguistic approach.

Damien Poirier, Cécile Bothorel, Émilie Guimier De Neef, Marc Boullé

France Telecom RD, TECH / EASY
2 avenue Pierre Marzin, 22300 Lannion, FRANCE
firstname.name@orange-ftgroup.com

Abstract

Community sites are by nature dedicated places to express and publish opinions. *www.flixster.com* is an example of participative web site, with dozens of millions of enthusiasts sharing their feelings/views on movies, providing positive feedback as well as vivid critics. For anyone interested in understanding net user expectations, such web sites are of major importance because they offer the opportunity to probe huge volume of user generated contents. But to actually benefit from those large amount of data, one has to be able to automatically extract users opinions. This is the challenge we tackle in this paper. Our goal is to exploit the various reviews written by a user in order to compute a model which can then be used to predict the user's verdict on a movie. We explore two different methods to extract opinions. The first one relies on a machine learning technique based on a naive bayesian classifier. The second method consists in applying NLP techniques to process opinions and build dictionaries : those dictionaries are then used to determine the polarity of a comment given the words it may contain. We did apply those two approaches to contents from flixster.com : the results we provide enable us to discern the most appropriate approach for a given set of data.

1. Introduction

With the spread of high speed access to the internet and new technologies, there is a tremendous growth in online music and video market. As more players appear on this field, competition increases and content provider can no longer wait for the customer. Instead they try to trigger purchases by pushing contents : suggesting different choices of movies or songs has become the big thing when it comes to sell content on-line. Actually recommendation is not a new concept, it is already used on internet commercial sites (*Amazon*, *Fnac*, *Virgin* ...) as well as on musical platforms (*Lastfm*, *Radioblog*, *Pandora* ...). But looking at the recommendation techniques used on such web sites shows there is room for innovation.

Candillier et al. (2007) presents an overview of recommendation techniques. These techniques are either based on internet users notations or content descriptions (*user-* and *item-based* techniques using collaborative filtering), or based on matching Internet user profiles and content descriptions (content filtering), or based on hybrid techniques combining both approaches. Although these techniques are different, they have the same problems: the hollow nature of matrix describing users and content profiles. Indeed, the sites proposing recommendations to their customers often have a large catalogue while users only give their opinion on a very low number of products. This phenomenon makes the comparisons between profiles risky. In the recommendation field, the difficulty to collect descriptions about users taste (rates, interests ...) and content (metadata) is a recurrent problem.

In order to compensate for these problems, a new research lane is open : mining the resources of the *open* Internet to boost *closed* sites performance. Instead of focusing solely on the data that can be retrieved from a single web site, recommendation techniques should shift to the vast amount of data that is now available from the Internet. In the era of Web 2.0 and community sites, it is now common for users to share pictures, tags, news, opinions ... Such data could be gathered to support automatic information extraction. Considering Internet like a wide open catalogue opens the way to learn the tastes of a large number of people : in the future, it could be possible to describe fan profiles, film typology or to discover new models to describe films and provide decisive pieces of advice on which films to recommend.

Motivated by this potential shift in recommendation, the purpose of this study is to extract opinions from movie reviews published on community sites[1]. Our main objective is to establish a user profile based on what he/she declares to like or dislike in movies through his/her published writings (blogs, forums, personal page on the flixster website ...).

We focus on two different approach to do so. The first method consists in applying a machine learning technique to classify textual reviews into either a positive or negative class. The second method consists in using a NLP approach to build an opinion dictionary and to detect words carrying opinion in the corpus and then predict an opinion.

We did apply those two approaches to data from the flixster web site. We discuss the results to compare the two approaches and we provide insights as to which approach should be used for a given corpus of opinions.

[1]This work enters in the frame of european project IST Pharos (PHAROS is an Integrated Project co-financed by the European Union under the Information Society Technologies Programme (6th Framework Programme), Strategic Objective "Search Engines for Audiovisual Content" (2.6.3))

2. Related work

Opinion extraction in trademark product reviews is a stake so important that a lot of researches have been done in the field. Dave et al. (2003) present a method for automatically classify reviews according to the polarity of the expressed opinions, i.e. the tool labels reviews positively or negatively. They index opinion words and establish a scale of rates according to intensity of words. They determine words intensity by using machine learning techniques. Finally, to classify a new review, they build an index reflecting the polarity of each sentence by counting identified words.
In an article by Morinaga et al. (2002), the authors explain how they verify reputation of targeted products by analyzing customers' opinions. They start by seeking Web pages *talking* about a product, for example a television, then they look for sentences which express opinions in these websites, and finally they determine if the opinions are negative or positive. They determine it by locating in reviews opinion words which were indexed previously in an *opinion dictionary*.
Other articles present works which are closely related to the previous one like Turney (2002), which classifies reviews in two categories: recommended and not recommended, or Wilson et al. (2004) which categorizes sentences according to polarity and strength of opinion, or Nasukawa and Yi (2003) which seeks opinions on precise subjects in documents.

We find two distinct types of methods: methods based on Natural Langage Processing (NLP) techniques and methods based on machine learning techniques. These two methods types can also be combined.

2.1. Linguistic methods of opinion analysis

Liu et al. (2005) describe a system which compares competitive products by using product reviews left by the Internet users. The system, named *Opinion Observer*, finds features such as pictures, battery, zoom size, etc. in order to explain the sentiment about digital cameras. They designed a supervised pattern discovery method to automatically identify the product features described in the reviews. A language pattern constrains a sequence of words and can be instanciated in many ways: *included/VERB [feature]/NOUN */VERB stingy/ADJECTIVE*. From the multiple instanciations, they extract association rules to find out what describes each feature: $noun1 noun2 \Rightarrow [feature]$.
They only keep the statistical relevant rules, and then generate language patterns: *noun1 [feature] noun2*. They analyse the reviews with those patterns and compare the opinion on each of these characteristics. A component decides the orientation of the extracted feature according to the words extracted near the features. Then they classify sentences as negative or positive by determining the dominant orientation of the opinion words of the sentence. The result of the comparison between two products is given in the form of diagram with features on X-coordinate and opinions polarity on Y-coordinate.

Opinion Observer is an example of a complete system based on the fine analysis of sentences and a process counting the Sentiment signs (words, expressions, patterns). Like many others (Morinaga et al., 2002; Turney, 2002; Wilson et al., 2004; Nasukawa and Yi, 2003), they need an *Opinion Dictionary* with as more words or expressions as possible expressing opinions. To build such a dictionary, different techniques are possible but they have all the same first steps : creating, manually, a set of words and expressions carrying opinion; this set is called *seed*; from the seed, the aim is to find other words and expressions yielding opinions and classify them according to their semantic orientation (positive, negative, but seldom neutral).

Lexicon can be built by using machine learning techniques. For example, Hatzivassiloglou and McKeown (1997) or Turney and Littman (2004) use an unsupervised learning algorithm to associate new words with words already registered. Pereira et al. (1994) and Lin (1998) describe methods to discover synonyms by analyzing words collocation.
Linguistic methods exploit syntactic and grammatical analyzis in order to extend the lexicon. Hatzivassiloglou and McKeown (1997) use conjonctions between a word which semantic orientation is known and a not classified word. For example, if there is the conjunction *and* between two adjectives, we can consider that the terms have a close signification. On the contrary, if there is the conjunction *but* between two adjectives, we can suppose that the two words have a different semantic orientation.

Turney (2002) uses a little more complex patterns. They count the frequency of the words or expressions beside a word or expression already classified and define the semantic orientation of those new words or expressions according to their neighbours. Each time they meet an adverb or an adjective, they extract a pair of consecutive words:

- Adjective with noun
- Adverb with adjective when they are not followed by a noun
- Adjectif with adjective when they are not followed by a noun
- Noun with adjective when they are not followed by a noun
- Adverb with verb

The second extracted word allows to confirm polarity of the adjective or adverb by giving an outline of the context of the sentence.
This method, counting co-occurences with words semantically oriented and manually selected, is also used in the research by Yu and Hatzivassiloglou (2003) in order to determine which words are semantically oriented, in which direction and the strength of their orientation. To measure more precisely the strength of opinion expressed in a sentence, a mean is to extract adverbs which are associated to adjectives. Indeed, Benamara et al. (2007) propose a classification of adverbs into five categories : adverbs of affirmation, adverbs of doubt, adverbs of weak intensity,

adverbs of strong intensity and adverbs which have a role of minimizer. A system of attribution of points according to the category of the adverb allows to calculate strengths to adverb-adjective combinations.

Google's work (Godbole et al., 2007) find semantic orientation of new words from WordNet databases (Miller et al., 1993). In a close manner, Hu and Liu (2004a) use sets of synonyms and antonyms present in WordNet to predict semantic orientation of adjectives. In WordNet, words are organised in tree (see figure 1). To determine polarity of a word, they traverse the trees of synonyms and antonyms of this word and if they find a seed word in the synonyms, they allocate the same class, but if they find seed word in the antonyms, they allocate the opposite class. If they do not find any seed word, they remake the analysis with synonyms and antonyms, and so on until finding a seed word.

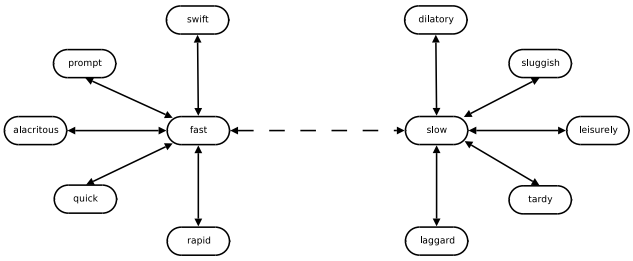

Figure 1: Tree of synonyms and antonyms in WordNet (full arrow = synonyms, dotted arrow = antonyms)

We think this method is a little too random because words can have different meaning according to the context and thus they can have synonyms not significating the same thing. For example, the word *like* has for synonym *love* but in the sentence *It is like that*, it has not the same meaning. This method finds a positive opinion in this sentence whereas there is not. But using the same method after having linguistically processed the corpus before, i.e. grammatical analysis, could be more effective. For the precedent example, if the seed word is *like/VERB*, we would not find opinion in the sentence *It is like that*.

To associate a polarity, negative or positive, to a sentence, we can count the number of terms with positive semantic orientation and the number of terms with negative semantic orientation. If there are more positive terms, the sentence is declared positive, if there are more negative terms, the sentence is declared negative, and if there are as many positive as negative terms, either sentence is declared neutral (Yu and Hatzivassiloglou, 2003); with another strategy, the last term carrying opinion determines the sentence polarity (Hu and Liu, 2004a). Otherwise, we can extract opinion one by one associated with the feature it refers to (Wilson et al., 2004; Hu and Liu, 2004b).

2.2. Machine learning for opinion analysis

Systems using learning machine techniques generally classify textual comments in two classes (positive and negative), but sometimes seek to predict five rates or more. These supervised classification methods consider that a comment describes only one product and try to predict the rate given by the author.

Many methods use NLP techniques to prepare the corpus. Wilson and Wiebe (2003) expose how to label opinion words with an intensity; Wilson et al. (2004) test three different learning methods, frequently used by the linguists: *BoosTexter* (Shapire and Singer, 2000), *Ripper* (Cohen, 1996) and *SVMlight*, the light version of Support Vector Machine by Joachims (1998). The last one obtains the best results on their annotated corpus. Pang et al. (2002) use a naive bayes classifier and a classifier maximizing the entropy. In the same way, in order to characterize what is appreciated or not in a sentence, Nigam and Hurst (2004) combine a *parsing* technique with a bayes classifier to associate polarity to sets of themes.

In addition, Pang et al. (2002) and Dave et al. (2003) show that corpus preparation with a lemmatizer or a negation detection for example, is useless. In order to predict reviews opinion, these two papers explore some learning methods and show that they are more powerful than *parsing* methods followed by a calculation as presented in the previous section. Considering comments as bags of words and using the relevant learning technique lead to 83% of good predictions. We will see in the following part of this paper that our own experiments confirm those conclusions.

3. Our two approaches

We compare in this section two opinion analysis approaches with their results. The initial corpus is composed of 60,000 films reviews rated by authors. Half of them express positive opinion and the other half, negative opinion. We keep a set of 10,000 positive and 10,000 negative for the tests. Both approaches are tested on the same test corpus.

The main difficulty of this corpus is the small size of reviews (twelve words on average). This makes opinion extraction difficult even for human sometimes. Moreover, the corpus is composed of textual messages very similar to forum messages. They present common characteristics such as accumulation of the ponctuation (" !!! "), smileys (" :-) "), SMS language (" ur ", " gr8 ") or words stretching (" veryyyyy cooooool ").

Each review in our corpus has a rate given by the author (0 to 5 stars) and our final aim is to predict this rate. We have decided to classify reviews in two classes. Reviews with a rate lower than three stars are considered as negative reviews, other as positive reviews. Here follow examples of reviews with their rate (table 1).

3.1. Linguistic approach

3.1.1. Technique

First step for this method is, as it was seen in the state of the art, the building of a dictionary of opinion words. We have used linguistic techniques to do that.

Rate	Review
POS	Great movie!
NEG	this wasn't really scary at all i liked it but just wasn't scary...
POS	I loved it it was awsome!
NEG	I didn't like how they cursed in it.....and this is suppose to be for little kids....
NEG	Sad ending really gay
POS	sooo awsome!! (he's soo hot)
POS	This is my future husband lol (orlando bloom)
NEG	Will Smith punches an alien in the face, wtf!!??
NEG	i think this is one of those movies you either love or hate, i hated it! :o)

Table 1: Examples of reviews

In first, we have separated all reviews according to their rate. For each review category (set of reviews rated 1 star, set of reviews rated 2 stars ...), we have applied a shallow parser (de Neef et al., 2002) to lemmatize and tag the text. We have filtered the words according to their Part of Speech tag and frequency. Verbs and adjectives have then been manually classified according to the opinion they convey.
This list has been increased using a synonym dictionary (*www.wordreference.com*). Only verbs and adjectives that are not ambiguous have been classified. For example, the word *terrible* is not classified because it can expressed both opinion polarities.

183 opinion words have been classified in two classes, positive words (115) and negative words (68), in this manner. The table 2 presents a part of the lexicon. Let us note this dictionary was not made on the corpus used to evaluate this method.

Positive words	good, great, funny, awesome, cool, brilliant, hilarious, favourite, well, hot, excellent, beautiful, fantastic, cute, sweet ...
Negative words	bad, stupid, fake, wrong, poor, ugly, silly, suck, atrocious, abominable, awful, lamentable, crappy, incompetent ...

Table 2: Part of hand crafted lexicon

The last step of the analysis consists in counting opinion words in each review to determine the polarity. For that we have in first time lemmatized all reviews (same pretreatment than in the lexicon building) and we have only kept adjectives and verbs. Then, we have assigned a polarity to reviews according to the majority number of positive words or negative words.
We have not performed any sophisticated NLP techniques such as a grammatical structural analysis. But keeping only verbs and adjectives avoid misinterpretations of words such as "*like*" which can hold different roles in a sentence. Regarding the review style, we can suppose that NLP tools would not face the bad English writing, and indeed, apply more complicated NLP treatments would probably became rapidly costly in adaptation to this specific corpus.

3.1.2. Results

This method allowed to rate 74% of films reviews on the 20,000 present in the test corpus. All the following results are calculated according to the rated reviews. To compare results with other techniques, we calculate three values: precision, recall and F_{score}.

Here follows the functions used to calculate these values:

- $precision = \frac{number\ of\ positive\ examples\ cover}{number\ of\ examples\ cover}$
- $recall = \frac{number\ of\ positive\ examples\ cover}{number\ of\ positive\ examples}$
- $F_{score} = \frac{2 * precision * recall}{precision + recall}$

The confusion matrix of results is presented in table 3.

	Pos. reviews	Neg. reviews
Predicted pos. reviews	8089	3682
Predicted neg. reviews	218	2823

Table 3: Confusion matrix obtained with the hand crafted lexicon

With this technique we obtain 0.81 for precision, 0.70 for recall and 0.75 for F_{score}.

The largest difficulty is to determine polarity of negative reviews. Indeed, the recall of negative reviews is 0.43, whereas it is 0.97 for positive reviews. Contrary, precision of positive reviews (0.69) is worse than precision of negative reviews (0.93).
This phenomon can be due to the dictionary we used: the positive category contains almost twice more words than negative category. But the problem is not the detection of negative reviews but their bad interpretation. These results lead us to think that people use negation to express their bad feelings sometimes without using any adjective nor verb carrying negative opinion. This intuition will be confirmed with the results of statistic approach.

To check quality of our dictionary, we have remade this experiment by using a English words set already classified by Stone et al. (1966) and Kelly and Stone (1975). The new lexicon contains 4,210 opinions words (2,293 negative words and 1,914 positive words). With this new opinion dictionary, the technique classifies more reviews (a gain of 4% essentially on negative ones) but results of prediction are worse than previously: 0.67 for precision, 0.65 for recall and 0.66 for F_{score}. See the confusion matrix of results in table 4.

The explanation for these worse results is certainly a lexicon less adapted for this corpus. It is a lexicon more general whereas our homemade lexicon was build with

	Pos. reviews	Neg. reviews
Predicted pos. reviews	7027	3743
Predicted neg. reviews	1165	3716

Table 4: Confusion matrix obtained with General Inquirer lexicon

words appearing regularly in a similar corpus.

These new results show the same problem with negative reviews, although this second lexicon contains more negative words. This confirms our first idea, negation is an important point to well interpret negative reviews.

3.1.3. Observation of the errors

Not rated reviews There are several explanations why reviews are not been rated:

- Gaps in the hand crafted lexicon. Examples: "wooohooo film", "watched it all the time when i was younger", "no please no", "I can not remember story".
- Presence of adjectives expressing sentiments or beliefs that can be associated with different opinion according to people. Examples: "so romantic", "weird movie", "I was afraid", "it is very sad".
- Presence of as many positive words as negative words. In this case, the classifior considers the review as neutral. Examples: "bad dish good opinion", "not bad - not great either", "really bad film, I thought it would be alot better".
- Some of the reviews get empty after NLP pretreatment. They are not containing any verbs nor adjectives.

Bad rated reviews Majority of errors are due to negation words which are not considered. The solution could be to change opinion polarity when a negation is present in the review. Indeed, reviews are very short so we can think that, statistically, they are composed of only one sentence, thus the negation modify on all the verbs or adjectives present. If this method is not satisfying, the idea could be to do a dependency parsing in order to find which word the negation is related to, and thus reversing the polarity only on the involved words.
We can find numbers of ironic or sarcastic sentences as "fun 4 little boys like action heros and stuff u can get into it :p" which was rated negatively by the author whereas we rate it positively.

3.2. Machine learning approach

Let us first present the method we used and then comment the results. We will analyze the prediction quality of our classifier, but we will show that a deeper exploration give information on the Internet users' writing style.

3.2.1. Compression-Based Averaging of Selective Naive Bayes Classifiers

In this section, we summarize the principles of the method used in the experiments. This method, introduced in Boullé (2007), extends the naive Bayes classifier owing to optimal preprocessing of the input data, to an efficient selection of the variables and to an averaging of the models.

Optimal discretization The naive Bayes classifier has proved to be very effective on many real data applications (Langley et al., 1992; Hand and Yu, 2001). It is based on the assumption that the variables are independent within each output label, and simply relies on the estimation of univariate conditional probabilities.
The evaluation of the probabilities for numeric variables has already been discussed in the literature (Dougherty et al., 1995; Liu et al., 2002). Experiments demonstrate that even a simple equal width discretization brings superior performance compared to the assumption using a Gaussian distribution.
In the MODL approach (Boullé, 2006), the discretization is turned into a model selection problem. First, a space of discretization models is defined. The parameters of a specific discretization are the number of intervals, the bounds of the intervals and the output frequencies in each interval. Then, a prior distribution is proposed on this model space. This prior exploits the hierarchy of the parameters: the number of intervals is first chosen, then the bounds of the intervals and finally the output frequencies. The choice is uniform at each stage of the hierarchy.
Finally, the multinomial distributions of the output values in each interval are assumed to be independent from each other. A Bayesian approach is applied to select the best discretization model, which is found by maximizing the probability $p(Model|Data)$ of the model given the data.
Owing to the definition of the model space and its prior distribution, the Bayes formula is applicable to derive an exact analytical criterion to evaluate the posterior probability of a discretization model.
Efficient search heuristics allow to build the most probable discretization given the data sample. Extensive comparative experiments report high performance.

Bayesian Approach for Variable Selection The naive independence assumption can harm the performance when violated. In order to better deal with highly correlated variables, the selective naive Bayes approach (Langley and Sage, 1994) exploits a wrapper approach (Kohavi and John, 1997) to select the subset of variables which optimizes the classification accuracy.
Although the selective naive Bayes approach performs quite well on datasets with a reasonable number of variables, it does not scale on very large datasets with hundreds of thousands of instances and thousands of variables, such as in marketing applications or, in our case, text mining. The problem comes both from the search algorithm, whose complexity is quadratic in the number of the variables, and from the selection process which is prone to overfitting.
In Boullé (2007), the overfitting problem is tackled by relying on a Bayesian approach, where the best model is found by maximizing the probability of the model given the data. The parameters of a variable selection model are the number of selected variables and the subset of variables. A hierarchic prior is considered, by first choosing the number of selected variables and second choosing the subset of se-

lected variables. The conditional likelihood of the models exploits the naive Bayes assumption, which directly provides the conditional probability of each label. This allows an exact calculation of the posterior probability of the models.
Efficient search heuristic with super-linear computation time are proposed, on the basis of greedy forward addition and backward elimination of variables.

Compression-Based Model averaging Model averaging has been successfully exploited in Bagging Breiman (1996) using multiple classifiers trained from re-sampled datasets. In this approach, the averaged classifier uses a voting rule to classify new instances. Unlike this approach, where each classifier has the same weight, the Bayesian Model Averaging (BMA) approach (Hoeting et al., 1999) weights the classifiers according to their posterior probability.
In the case of the selective naive Bayes classifier, an inspection of the optimized models reveals that their posterior distribution is so sharply peaked that averaging them according to the BMA approach almost reduces to the MAP model. In this situation, averaging is useless.
In order to find a trade-off between equal weights as in bagging and extremely unbalanced weights as in the BMA approach, a logarithmic smoothing of the posterior distribution called compression-based model averaging (CMA) is introduced in Boullé (2007).
Extensive experiments have demonstrated that the resulting compression-based model averaging scheme clearly outperforms the Bayesian model averaging scheme.

3.2.2. Results

With this approach we have no *a priori* on the data. Indeed we hang on all reviews as the authors wrote them and process them as bags of words. We do not treat the data with NLP tool. We apply to the text only two treatments; we put in lowercase all letters and we delete the punctuation.

We learned on a corpus containing 20,000 positive reviews and 20,000 negative. We tested this training on the same test corpus than in the precedent method.

Let us start by commenting training results. The tool found 305 informative variables out of the 24,825 words present in the learning corpus. Little of them are very informative as shown in the figure 2. They are classified according their *level* value. The *level* is directly related to the posterior probability of a discretization model, with a 0-1 normalization. Its value is 0 in case a no informative input variable and is asymptotically equal to 1 in case of perfectly informative input variable.
Majority of words having a positive level express opinion. But other words appear in this list.

These results allow to learn opinion vocabulary but also information on the style of the reviews. Informations supplied by "*and*" (table 5) indicate that authors write longer texts with more details when they talk about a movie they appreciated.
This phenomenon is specified with other terms present in the list. We find too "*movie*" (table 6) and "*film*" (table 7)

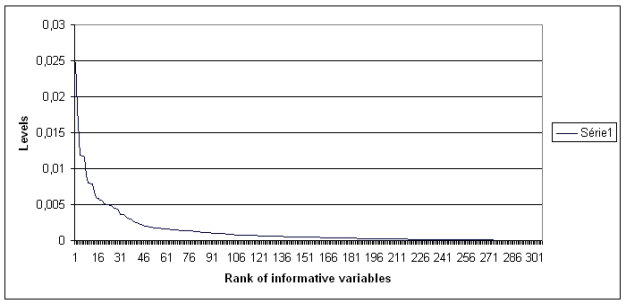

Figure 2: Evolution of levels of informative variables

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,525397	0,474603	33725
[0.5; 1.5[	0,401667	0,598333	4439
[1.5; 4.5[	0,299568	0,700432	1619
[4.5; inf[	0,0599078	0,940092	217

Table 5: Informations of "*and*"

and link-words as "*a*" (table 8), "*the*" (table 9), "*of*", "*in*" ...
One can think that users have tendency to be more prolix and detail their point of view on film features when they appreciate the movie.
The presence of words as "*action*" (table 10) and "*thriller*" (table 11) can confirm this explanation. Authors explain why they appreciated the film and what they appreciated in the film.

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,535447	0,464553	30849
[0.5; 1.5[	0,380505	0,619495	9151

Table 6: Informations of "*movie*"

Another observation is the presence of negation words in informative variables. That explain certainly the weak score of precision for positive reviews and recall for negative reviews in the previously approach. Indeed, we can note that negation terms appear much more in negative reviews than in positive reviews (table 12 and 13).

Concerning the opinion prediction, the confusion matrix of results in table 14 shows that this time, all the reviews are classified. Scores obtained are 0.77 for precision, 0.76 for recall and F_{score}. They are better than those obtained with the classic naive Bayes classifier (approximately 0.70 for the three indicators). Results are equivalent to our linguistic results regarding to the F_{score}, but, recall is significantly better for negative reviews (0.82 instead of 0.43), also is the precision on positive reviews (0.80 instead of 0.69). On the contrary, recall is worse for positive reviews (0.70 instead of 0.97) and so is the precision on negative reviews (0.74 instead of 0.93). ML technique provides balanced results for each class, but overall it does not outperforms the NLP approach.

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,513252	0,486748	37013
[0.5; 1.5[	0,335788	0,664212	2987

Table 7: Informations of "*film*"

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,523142	0,476858	31177
[0.5; 1.5[	0,430528	0,569472	7960
[1.5; 2.5[	0,304751	0,695249	863

Table 8: Informations of "*a*"

4. Conclusion, prospects

We have tested and evaluated two approaches for opinion extraction. The first one consists in building a lexicon containing opinion words using *low-level* NLP techniques. This lexicon allows to classify reviews as positive or negative. The second method consists in using a machine learning technique to predict the polarity of each review.

We used data from flixster as a benchmark to evaluate those two recommendation methods, using part of the opinion corpus as a learning testbed and the rest of it to evaluate classification performance. Thanks to those experiments, we are able to discriminate the qualities of the two techniques according to various criteria. In the rest of this conclusion, we synthesize our results, trying to provide the reader with an understanding of each technique specificity and limitation.

While digging into the results obtained with the machine learning (ML) technique, it seems that it inherently provides a deeper understanding of how the authors express themselve according to what they thought about a movie. Indeed, results show that people generally write more when they appreciated the movie for example, giving more detailled reviews of movies features. It turns out that opinion words are not the only opinion indicator, at least for this kind of corpus.

Independently of the analysis technique, an important issue with automating opinion extraction is that we cannot expect a machine to predict good polarity for each review. Consider for instance the sentence "*Di Caprio is my future husband*": it does not indicate whether the author appreciated the film or not. Thus our aim is not to know the polarity of each review but to have the best possible classification (including indetermination). Improvement of prediction results with ML can be obtained by using an indecision threshold. i.e. when the probability to have a well prediction is too weak, we can decide not to classify the review.

With NLP technique, this problem does not exists because reviews which do not contain opinion words are not classified. However, results of this technique can be improved. For instance detectinng negations would be an important progress. Indeed, ML results show that negative opinions are often expressed by using words carrying positive opinion associated with a negation. Since our linguistic approach ignores every negations, most of the negative reviews are labelled as positive ones. Best solution is probably to proceed to a dependency parsing. But the kind of prose which we are faced with (SMS writing, spelling errors, weird sentences construction . . .) certainly will complicate this step.

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,522945	0,477055	29179
[0.5; 1.5[	0,457694	0,542306	8923
[1.5; 2.5[	0,346154	0,653846	1898

Table 9: Informations of "*the*"

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,505069	0,494931	39262
[0.5; 1.5[	0,230352	0,769648	738

Table 10: Informations of "*action*"

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,502907	0,497093	39725
[0.5; 1.5[	0,08	0,92	275

Table 11: Informations of "*thriller*"

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,48512	0,51488	36189
[0.5; 1.5[	0,641301	0,358699	3811

Table 12: Informations of "*not*"

Value	Neg. review	Pos. review	Frequency
]-inf; 0.5[	0,49419	0,50581	38896
[0.5; 1.5[	0,70471	0,29529	1104

Table 13: Informations of "*didn't*"

	Pos. reviews	Neg. reviews
Pos. reviews predict	7060	1793
Neg. reviews predict	2940	8207

Table 14: Confusion matrix obtained with Machine Learning

The main point characterising ML techniques is that new datasets can be analysed without any *a priori* knowledge (i.e. lexicon) and then quickly deployed with a confortable reliability on both positive and negative reviews. But the corpus has to be large enough to offer a consistent training dataset and has to contain rates to supervise the training, which is not always the case.

This approach may also be used to detect pertinent words and thus help in building the dictionary, particularly in the context of Web Opinion Mining, where it is necessary to adapt the lexicon to the *inventive* vocabulary the Internet users' writings abound in.

Contrary, NLP technique does not require learning step, except regular updates of the lexicon. So it can be deployed immediately on a small corpus without rated examples. With a dependency parsing step in order to detect negations, the results could be competitive with ML techniques if not more often.

As a conclusion, we propose to use a *low-level* NLP approach when the corpus is too small to have a good training: the cost of building a lexicon (small ones bring satisfying quality) and designing a negation detection remains reasonable. If the corpus is large enough, ML approach will be easier to deploy.

To go further, we may explore if linguistic pretreatments on the corpus for ML technique can reduce the number of variables (by reducing the vocabulary describing the reviews) without losing information and damaging the quality. We may also focus on a higher level NLP approach and try to explain why people (dis)like movies.

5. References

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and VS Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better then adjectives alone.

M. Boullé. 2006. MODL: a Bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1):131–165.

M. Boullé. 2007. Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research*, 8:1659–1685.

L. Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Laurent Candillier, Frank Meyer, and Marc Boullé. 2007. Comparing state-of-the-art collaborative filtering systems. International Conference on Machine Learning and Data Mining MLDM 2007, Leipzig/Germany.

William W. Cohen. 1996. Learning trees and rules with set-valued features.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.

E. Guimier de Neef, M. Boualem, C. Chardenon, P. Filoche, and J. Vinesse. 2002. Natural language processing software tools and linguistic data developped by france télécom rd. Indo European Conference on Multilingual Technologies, Pune, India.

J. Dougherty, R. Kohavi, and M. Sahami. 1995. Supervised and unsupervised discretization of continuous features. In *Proceedings of the 12th International Conference on Machine Learning*, pages 194–202. Morgan Kaufmann, San Francisco, CA.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. ICWSM'2007 Boulder, Colorado, USA.

D.J. Hand and K. Yu. 2001. Idiot bayes ? not so stupid after all? *International Statistical Review*, 69(3):385–399.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives.

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. 1999. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews.

Minqing Hu and Bing Liu. 2004b. Mining opinion features in customer reviews.

Thorsten Joachims. 1998. Making large-scale support vector machine learning practical.

Edward Kelly and Philip Stone. 1975. Computer recognition of english word senses.

R. Kohavi and G. John. 1997. Wrappers for feature selection. *Artificial Intelligence*, 97(1-2):273–324.

P. Langley and S. Sage. 1994. Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 399–406. Morgan Kaufmann.

P. Langley, W. Iba, and K. Thompson. 1992. An analysis of Bayesian classifiers. In *10th national conference on Artificial Intelligence*, pages 223–228. AAAI Press.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words.

H. Liu, F. Hussain, C.L. Tan, and M. Dash. 2002. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 4(6):393–423.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Introduction to wordnet: An on-line lexical database.

Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. 2002. Mining product reputations on the web.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing.

Kamal Nigam and Matthew Hurst. 2004. Towards a robust metric of opinion.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1994. Distributional clustering of english words.

Robert E. Shapire and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. 1966. The general inquirer: A computer approach to content analysis.

Peter D. Turney and Michael L. Littman. 2004. Unsupervised learning of semantic orientation from a hundred-billion-word corpus.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews.

Theresa Wilson and Janyce Wiebe. 2003. Annotating opinions in the world press.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? finding strong and weak opinion clauses.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences.

Co-Word Analysis for Assessing Consumer Associations: A Case Study in Market Research

Torsten Teichert[1], Gerhard Heyer[2], Katja Schöntag[1], Patrick Mairif[2]
[1]University of Hamburg, Marketing and Innovation
Von-Melle-Park 5, D-20146 Hamburg
[2]University of Leipzig, Natural Language Processing
Johannisgasse 26, D-04103 Leipzig
E-mail: teichert@econ.uni-hamburg.de, heyer@informatik.uni-leipzig.de, schoentag@econ.uni-hamburg.de, pmairif@informatik.uni-leipzig.de

Abstract

Sentiment analysis is particularly relevant in marketing contexts because it is essential for deriving an in-depth understanding of consumer behaviour. This manuscript illustrates an exemplary best-practice case study for the application of text analysis tools. The case study analyzes the association of female consumers with the product category "shoes". Automated text analysis is used to identify features and structures from the qualitative data at hand. The results of the automated text analysis are contrasted with manual feature coding, showing a comparable coding quality while yielding considerable savings of time and effort. Thus we conclude that NLP offers a high potential for future research applications to solve marketing problems.

1. Introduction

With the advancement of natural language processing (NLP), many new research opportunities have opened up for scientists in various different fields. Among these, marketing research constitutes a prominent but yet unexplored application field. Recent innovative approaches in this area rely on in-depth interviewing to gain insight into consumers' thoughts and feelings regarding specific brands and products (Teichert et al., 2004). Interviews yield a large amount of qualitative data that is hard to handle and needs to be structured in order to be analyzed. Manual coding and categorization can be cumbersome and time consuming. Therefore, the development and application of text analysis software is of high importance for marketing researchers both on the academic and the practitioners' side.

This manuscript illustrates an exemplary best-practice case study for the application of text analysis tools. Sentiment analysis is particularly relevant in marketing contexts because it is essential for deriving an in-depth understanding of consumer behaviour. This case study analyzes the association of female consumers with the product category "shoes". This product category is assumed to be emotionally laden especially for female consumers, reaching far beyond mere functional aspects. Data elicitation and processing techniques are based on methods derived from human associative memory models and network analysis. As opposed to many text analysis applications, data are not obtained from secondary (internet) sources but from 30 personal in-depth interviews with female consumers. The underlying objective for the pursued marketing research is to derive a novel characterization of female shoe consumers. This should build the basis for developing innovative marketing measures which target yet unexplored consumer sentiments.

Automated text analysis is used to identify features and structures from the qualitative data at hand. Text analysis tools are integrated into qualitative data analysis in order to minimize subjectivity and to maximize replicability by excluding all elements of personal experience and emotions from the coding process. The results of the automated text analysis are contrasted with manual feature coding, showing a comparable coding quality while yielding considerable savings of time and effort. Thus we conclude that NLP offers a high potential for future research applications to solve marketing problems.

2. Conceptual Background

The specific requirements for text analysis tools are derived from theories and techniques underlying the applied elicitation and analysis techniques. It is essential to understand the working of consumer memory as well as the encountered problems in eliciting and analyzing qualitative data, in order to develop an appropriate procedure of automated text analysis.

2.1 Consumer Associations and Mental Processing

Various theories and models exist that explain the working of human memory. Particularly in the field of marketing and consumer behaviour research, Human Associative Memory (HAM) is a widely accepted model with an increasing number of studies

based upon it (Krishnan, 1996; Henderson et al., 1998; Henderson et al., 2002). According to this model, information is stored in nodes which are linked (associated) with each other forming a complex network of associations (Anderson, 1983; Keller, 1993). Based upon this, Spreading Activation theory provides a (however not uncriticized) framework to explain temporal aspects of associations (Hermann et al, 1993). It assumes that mental activity spreads from active concepts to all related concepts.

In the case of brands, for instance, the stimulating element can be a brand's logo or advertising jingle: individual nodes within the brand's associative network are activated and become accessible and retrievable. Activation then spreads to adjacent nodes turning activated nodes into source nodes which, in turn, spread their activation to their neighbour nodes (Anderson, 1983; Collins & Loftus, 1975). This spread of activation produces a chain, or flow, of thoughts. A representation of this flow of thoughts, though inevitably incomplete, can be obtained from the flow of speech, for example when eliciting brand or product associations during an interview. Speech not only contains the main aspects that are stored for a particular concept, i.e. the informational content of nodes, it can also be used to track the flow of thought and thus the existing associations, i.e. links between nodes, in the interviewee's mind.

Since most information is stored non-verbally in the human mind, standardized questionnaires and straightforward questioning often do not produce the desired results. Elicitation techniques, such as the Zaltman Metaphor Elicitation Technique (Zaltman & Coulter, 1995), take the non-verbal nature of human knowledge into account and aim at surfacing primary and secondary associations. Applying visual, projective, and sensory techniques helps access subconscious memory of episodic, autobiographic, visual and sensory nature as well as a metaphoric description of thoughts, sentiments, and emotions (Teichert et al., 2004).

2.2 Drawbacks of Manual Data Analysis

The scientific discourse reveals several basic problems of qualitative data analysis. When analyzing the flow of speech in the form of transcribed interviews, researchers are frequently faced with ambiguity of statements and expressions. In order to structure and code the text on hand, they are often required to interpret the interviewees' statements so that a rather subjective representation of the elicited data results. The replicability of the results is consequently rather low. An existing and widely used remedy to this problem is the parallel data processing and coding by two or more researchers.

The higher the inter-rater-reliability, i.e. the number of identical codes among the raters given independently of each other, the higher the objectivity level is assumed to be. By convention, inter-rater-reliabilities of 70 percent and above are acceptable. However, when there are differences between raters, codes are often assigned based on discussions which constitute another subjectivity factor. Particularly in the case of sentiment coding, it can be hypothesized that inter-rater-reliability is comparatively low for emotional aspects as opposed to more rational expressions.

Text analysis tools offer a solution to this problem as they reduce the level of subjectivity to a minimum, both during the feature extraction and the categorization processes. This leads to a high replicability level and, thus, to a higher level of reliability.

2.3 Requirements for Automated Co-Word Analysis

A tool had to be developed that addresses the particular requirements of sentiment analysis based on qualitative interviews with real-life consumers. The concept of Human Associative Memory guides the data processing and evaluation process by four main assumptions:

1. Words or concepts mentioned together are linked in the mind.
2. The more salient a concept is, the more often it is mentioned during the course of an interview.
3. The stronger the association between two concepts, the more often they are mentioned together.
4. Valence of a concept is indicated by positive or negative adjectives annotated to it.

Qualitative interview data consist of lengthy, often quite unfocused text information. Thus, text analysis tools first and foremost need to identify and extract the main consumer thoughts and sentiments from the transcribed interviews while excluding irrelevant aspects. Further, in order to make the sentiment data more manageable and interpretable, individual sentiments must be assigned to categories that represent specific types of feelings toward the product or brand in question. Finally, data needed to be coded such that network analytic evaluation techniques could be applied.

In sum, the specific requirements of the text analysis tool were as follows:

1. Extraction of features and consolidation of extracted features into meaningful categories.
2. Processing of the data using a co-word-analysis on a paragraph level as basis for the development of associative networks.
3. Consideration of valence expressions for the weighting of individual features.

The developed text analysis tool fulfils all of the above mentioned criteria while providing an intuitive user interface for marketing researchers without extensive IT background.

3. Technique and Implementation

The complete process covered by the tool comprises the following steps:

1. Import of text sources.
2. Processing of texts.
3. Graph creation
4. Graph clustering.

If there is a large amount of text to be processed, all processing steps can be done by an automatic batch process. To play with the data and tune parameters, the user is provided with a graphical user interface that reflects the processing structure and allows them to interfere at each processing step (cf. figure 1).

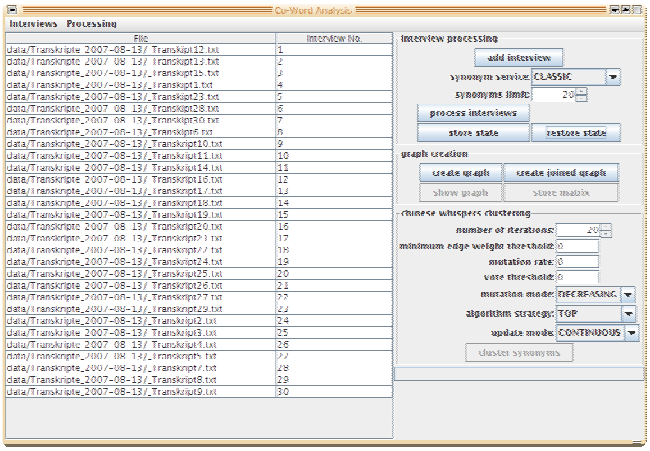

Figure 1: Graphical user interface

3.1 Import of Text Sources

The first task covers the reading of text files with annotated metadata in a specific file format. Texts are automatically divided into sections and paragraphs that contain answers to questions as part of interviews.

3.2 Processing of Text

The second task does the real work: the processing of texts. The aim is to extract features for sections and paragraphs, to reduce word forms to base forms and to add synonyms to each of them.
The extraction of features is mainly pattern based. First, we eliminate stop words and recognize valence words. Features are then supposed to be located right of valence words. Stop words are the usual high frequency general language and domain specific uninteresting words with the exception of valence markers. These are words that modify the meaning of the features. E.g. in a sentence such as “These shoes are very comfortable.”, the words “these” and “are” are recognized as stop words, “very” is a valence word and “shoes” and “comfortable” are features; “very” is associated to and located left of “comfortable”.

Each valence word has an associated value that modifies the value of the feature. These values can be positive or negative, e.g. “don't” would be considered a valence word with negative value. Each sentence is processed separately in order to avoid side effects with valence words at the end of a sentence. The text is searched for known valence words first (to ensure that we do not eliminate valence words if they appear in the list of stop words). Next, stop words are eliminated. What remains are the features and possibly valence words associated to them.

The resulting features are reduced to their base forms by using a web service for base form reduction offered by the department for natural language processing at the University of Leipzig (http://wortschatz.uni-leipzig.de/Webservices/). To get the base form of a given word, the web service simply uses a pre-processed list of words that associates with each word form a corresponding base form (Biemann et. al., 2004).

The base forms of the features are then used to request synonyms by the Leipzig web services. As a result, a list of weighted features is derived with a list of synonyms associated to each of them.
To make the data visible, HTML documents are generated containing the original texts and highlighting the extracted information (cf. figure 2). The features are marked in green or red with colour shades from dark to light green for positive and red for negative values; valence words are printed in italics; the base forms are added after the feature in square brackets and synonyms are shown as tool tips.

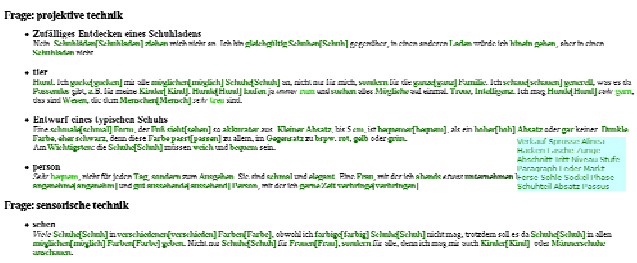

Figure 2: HTML visualisation of processed text

3.3 Graph Creation and Clustering

In order to cluster the features, we build a graph that associates similar features. To this end, the synonym vectors of the features are compared and a similarity value is calculated by comparing the synonym vectors with respect to the number of common synonyms using the Dice coefficient (Heyer et. al. 2006). Alternatively hereto, a cosine co-efficient could have been applied which showed to be more sensitive to the relative importance of a word (Lewis et al. 2006).

In addition to the creation of the "normal" graph, it is possible to create a "joined" graph in which nodes with a high[1] similarity are joined. In case there are too many nodes in the graph that have many links to each other, this joining of nodes generally achieves better clustering results.
The tool has functions to display the graph and to export it as a matrix. These functions do not change the derived results; they merely allow for a simple visualisation and make it possible to analyse the data with other tools.

The resulting graph is clustered with the Chinese Whispers Clustering algorithm (Biemann, 2006). Each feature is assigned to a cluster. The resulting data are then exported to a CSV[2] file that can easily be imported into other tools for further processing.

4. Exemplary Case Study

In order to illustrate the proposed approach, a study was conducted that analyzed the association of female consumers with the product category "shoes". This product category is assumed to be emotionally laden especially for female consumers, reaching far beyond mere functional aspects. While the comfort of shoes influences the physical well-being of a person, it is more than the mere satisfaction of such physical needs that drives the purchase of shoes. During the buying process, both physical products as well as brand images provide cues for the activation of associative networks, leading to the purchase or rejection of particular products. Using sentiment analysis to reveal subconscious associations helps marketers understand how consumers really feel about the product category and enables them to create effective targeted marketing programs.

The study was conducted in Hamburg, Germany, with 30 women between 23 and 57 years of age elaborating on their perception of shoes in 30-minute interviews each. The sample contained a mixture of women with various cultural, educational, and sociodemographic backgrounds. Further, as suggested by Supphellen (2000), the sample contained heavy users as well as average and light users. Several different questioning techniques, including the presentation of visual stimuli as well as sensory and projective techniques (Teichert et al, 2004) were applied. This allowed for a comprehensive view on interviewees' associations regarding shoes and the process of purchasing and wearing shoes.

A total of 1,938 different features could be extracted from the transcribed interviews. Manual coding resulted in 133 and 112 categories for the two raters respectively. Inter-rater-reliability was 65.3 percent after the exclusion of uncategorized features. Interestingly, inter-rater-reliability was 60.6 percent for emotional aspects while for rational aspects, it was 66.7 percent. Thus, the hypothesis of the increased subjectivity in the coding of sentiments could be confirmed.

The automatic categorization resulted in 185 categories or clusters. 100 of the 148 manually developed categories, i.e. 67.6 percent, were identical or similar to the automatically developed categories. This figure highlights the high quality and accuracy of the clustering algorithm.

The network analytic examination of the processed data yields an associative network as shown in figure 3. In order to reduce the complexity and make the network intuitively understandable, only links of strength 7 and above and the respective nodes are shown.

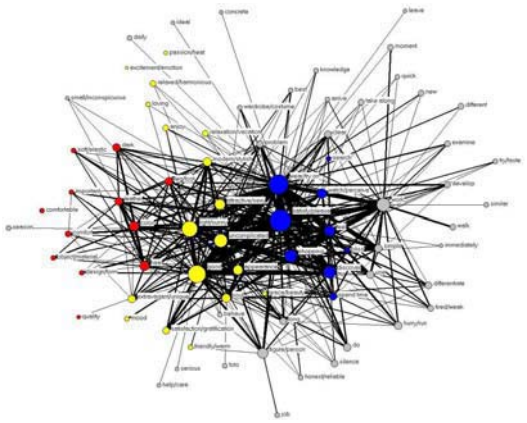

Figure 3: Associative Network for Shoes

It can clearly be seen that the product category of shoes activates a number of highly emotional associations in the female consumers' minds. The experiential aspects of the purchasing process (marked in blue) are highly salient as shown by associations such as "satisfy/please", "wear/try on", "spend time", "discover", "examine", "watch/perceive", "satisfaction/gratification", "enjoy", and "bliss." Simply put: the process of selecting and buying shoes makes female consumers happy and gives them a feeling of deep satisfaction. Service quality and store ambience can therefore be strong differentiating factors for a shoe or shoe store brand.

Additionally, shoes are not only seen as part of a woman's appearance, but they also contribute to her overall well-being and self-confidence. Strong links between associations such as "appearance", "attractive/sexy", "comfort", "grace/beauty", "extravagant/unique", "soul", and "mood" (marked in yellow) highlight the role of shoes as transformers of a woman's perception of herself. The associative network further reveals the noteworthy effect of high heels on a woman's way of walking: with the felt "extension" of her legs, the interviewed female consumers perceive to walk more gracefully and feel more attractive when wearing high heeled shoes.

[1] If the similarity exceeds a given value.
[2] Comma-separated values - http://en.wikipedia.org/wiki/Comma-separated_values

A third group of nodes (marked in red) comprises aspects of the actual quality of a shoe, including its design, form, material, and colour, which is part of a shoe's signalling function. Finally, associations such as "light/sunny", "uncomplicated" and "relaxed/harmonious" help researchers understand how women characterize their relationship with shoes: unlike clothing, shoes fit no matter whether a woman's weight changes, making them very "uncomplicated companions" and the shopping experience very "relaxed."

Translated into marketing activities, shoe brands could gain a significant competitive edge by using a strongly personal and emotional positioning. A communication strategy should be designed that reflects their remarkable transformative effect. Shoes are a highly personal issue, comparable to jewelry, which is why approaches such as mass customization using online design platforms that take a woman's desire for unique shoes into account may have a high potential for future success.

5. Conclusion and Outlook

As shown by the case study, automated text analysis offers many interesting opportunities for innovative marketing research applications. The developed tool yields results that are comparable to manual coding of qualitative data while requiring only a fraction of the necessary time and effort. The network representation of the main concepts offers a quick yet comprehensive overview of the complete pool of qualitative data. Further network analytic measures can yield more detailed insights into the roles and relationships of individual associations. Such consecutive analyses, which are beyond the scope of this article, should allow for additional concept disambiguation and a thorough analysis of the interviewees' thoughts and sentiments. This would hardly be possible with manual data processing and purely verbal descriptions of the findings.

Future work can improve the results in various ways. In the field of text analysis, the number of one-element clusters can be reduced. There are clustering algorithms that may perform less powerfully while possibly yielding better results (e. g. agglomerative hierarchical clustering). Also the synonym data is still incomplete. Thus, as of now, some features cannot be clustered because there are no synonyms to be compared. Another task is the extraction of valence words. At present, they are recognized only if they appear in front of features. But human language is more variable. For example consider the following sentence: "They may say it is delicious, but it is not!" To handle wording like this appropriately, more complex patterns will be needed.

On the marketing side, we see a range of opportunities arising from the application of automated text analysis. However, marketing researchers must look beyond the mere extraction and clustering of features. Taking the network structure of human memory into account, possible future research should aim at reconstructing both the order as well as direction of node activation. This would allow to gain an even deeper understanding of purchasing motivation and decision processes. Additionally, data elicitation and interview transcription techniques should be adjusted to ex-ante accommodate for the specifics of automated text analysis tools in order to yield the most useful and precise data possible.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press

Biemann, C., Bordag, S., Quasthoff, U., and Wolff, C. (2004). *Web services for language resources and language technology applications*. In: Proceedings Fourth International Conference on Language Resources and Evaluation, LREC 2004.

Biemann, C. (2006). *Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems*. Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06, New York, USA.

Collins, A. M., & Loftus, E. F. (1975). *A Spreading-Activation Theory of Semantic Processing*. Psychological Review, 82(6), 407-428.

Henderson, G. R., Iacobucci, D., & Calder, B. J. (2002). *Using Network Analysis to Understand Brands*. Advances in Consumer Research, 29(1), 397-405.

Heyer, G., Quasthoff, U., Wittig, T. (2006), *Text Mining – Wissensrohstoff Text*, W3L-Verlag, Bochum 2006.

Keller, K. L. (1993). *Conceptualizing, Measuring, and Managing Customer-Based Brand Equity*. Journal of Marketing, 57(1), 1-22.

Krishnan, S. H. (1996). *Characteristics of memory associations: A consumer-based brand equity perspective*. International Journal of Research in Marketing, 13(4), 389-405.

Supphellen, M. (2000). *Understanding core brand equity: guidelines for in-depth elicitation of brand associations*. International Journal of Marketing Research, 42(3), 319-338.

Teichert, T., Valta, K., von Weissenfluh, D. (2004). *Entwicklung einer "Marke Bern",* in: Zerres, Ch. (edt.), Markenforschung, Reiner Hampp, München, 311-329.

Zaltman, G., & Coulter, R. H. (1995). *Seeing the Voice of the Customer: Metaphor-Based Advertising Research*. Journal of Advertising Research, 35(4), 35-51.

Affect Transfer by Metaphor for an Intelligent Conversational Agent

Alan Wallington, Rodrigo Agerri, John Barnden, Mark Lee, Tim Rumbell
School of Computer Science
University of Birmingham,
Birmingham B152TT
E-mail: A.M.Wallington@cs.bham.ac.uk

Abstract

We discuss an aspect of an affect-detection system used in edrama by intelligent conversational agents, namely affective interpretation of limited sorts of metaphorical utterance. Our system currently only deals with cases, which we found to be quite common in edrama, in which a person is compared to, or stated to be, something non-human such as an animal, object, artefact or supernatural being. Our approach permits a limited degree of variability and extension of these metaphors We discuss how these metaphorical utterances are recognized, how they are analysed and their affective content determined and in particular how the Electronic Lexical Database, WordNet, and the natural language glosses of the WordNet sysnsets can be used. We also discuss how this relatively shallow approach relates in important ways to the deeper ATT-Meta theory of metaphor interpretation and to approaches to affect and emotion in metaphor theory. We finish by illustrating the approach with a number of 'worked examples'.

1. Introduction

This paper discusses aspects of the extraction and processing of affective information such as emotions/ moods (e.g. embarrassment, hostility) and particularly evaluations (of goodness, importance, etc.) as conveyed by metaphor in free-form textual utterances. The background to this work is our experience in building upon an edrama system produced by one of our industrial partners, in which human users - school children, so far, in the testing and development stage of our work- improvise around certain themes by typing in utterances for the on-screen characters they play to utter (via speech bubbles). Drama by its nature involves emotional experience and this is furthered by the nature of the themes or scenarios we have used, namely 'school bullying' and a scenario involving a sufferer of a particularly embarrassing disease -Crohn's disease- discussing with friends and family whether or not to undergo an operation. User-testing (Zhang et al, 2006) shows that users have enjoyed using the system.

The need for the extraction and processing of affect arises because we have added to the edrama the option of having a bit-part character controlled by an Intelligent Conversational Agent (ICA). This ICA is capable of making largely contentless, but emotionally appropriate, interjections and responses in order to keep the conversation flowing, which it does by extracting affect from the human controlled characters' utterances. The same algorithms are also used for influencing the characters' gesturing (when a 3D animation mode produced by one of our industrial partners is used). Whilst other ICA research has concerned itself with the conveyance of affect (e.g. Picard, 2000), it appears that the conveyance of affect via metaphor has been largely ignored. Indeed, relatively little work has been done on any detailed computational processing of metaphor. Major exceptions include (Fass, 1997; Hobbs, 1990; Martin, 1990; Mason, 2004; Narayanan, 1999).

The background to the work on the conveyance of affect via metaphor comes from the authors' approach to, and partially implemented system (ATT-Meta) for, the processing and understanding of metaphor in general (Agerri et al. 2007; Barnden et al. 2004; Wallington et al. 2006). This is a more ambitious aim than the mere recognition of a metaphor or the classification of a metaphor into one of a number of different metaphor classes or conceptual metaphors (see Mason, 2004). The details of the implemented system need not concern us since they are not used in the control of the edrama ICA. However, aspects of the approach to metaphor are used. Thus, our metaphor approach and system emphasizes the open-endedness of metaphorical expressions, whereby conventional metaphors and fixed phraseology may be varied, extended and elaborated upon so as to convey further information and connotations not conveyed by the conventional metaphor. Although our ICA work uses WordNet for analysis of many of the affect-conveying metaphorical senses we find, we can analyse some phrasal variation in the words and deal with some senses that are not found.

Relatedly, our approach and system eschews large sets of correspondences between ontologically complex source and target domains in the manner of Lakoff and Johnson's (1980) 'Conceptual Metaphor Theory' e.g. ARGUMENT IS WAR, or ANGER IS HOT LIQUID UNDER PRESSURE (see Gibbs, 1992; Kövecses, 2000), with the meaning of a metaphorical utterance 'read off' from the source-target correspondences. Instead we assume very few, more abstract, specific source-target links between domains and account for much of the apparent systematic relatedness between source and target domains by noting that certain types of information, relations, attributes that can be inferred as holding of the situation described by a metaphorical utterance transfer in an invariant manner to

the target via a limited number of what we term View-Neutral Mapping Adjuncts or VNMAs. For example, we assume that if a causal link can be inferred as holding between entities in the source, then the causal link will hold by default in the target. Similarly, if something applies to a particular degree in the source, then its target equivalent will apply to the same degree in the source and likewise with such information as duration, temporal ordering, logical relations between entities, and others. Crucially for our edrama ICA, we assume that the emotional state that is either invoked by some aspect of the source, or that holds within the source will carry over to the target. We also assume that a value judgement concerning something in the source will also carry over by default to the target. For example consider a situation in which it is said of some foul mouthed character, 'Tom is a sewer'. This can be partially analysed in terms of Reddy's (1979/1993) well known 'conduit metaphor', in which information and utterances are viewed as if passing along a conduit from speaker to hearer, but crucially no source-target correspondence will be required for the specifics of 'sewer'. Instead, the negative value judgement about the nature of the material passing through a sewer should be transferred by the Value Judgement VNMA. A similar negative value judgement is conveyed by 'smelly attitude' or by the comment 'you buy your clothes at the rag market', two examples taken from transcripts the system automatically recorded during user-testing.

Emotional states and behaviour are often described metaphorically (Kövecses, 2000; Fussell & Moss, 1998), as in 'He was boiling inside' [feelings of anger] or 'He was hungry for her' [feelings of lust] and conceptual metaphors such as the above mentioned ANGER IS HOT LIQUID UNDER PRESSURE or LUST IS HUNGER proposed to account for this, but in an analysis of the transcripts from our user-testing the type of affect laden metaphor described in the previous paragraph was found to be a significant issue in edrama: at a conservative estimate, at least one in every 16 speech-turns has contained such metaphor (each turn is 100 characters, and rarely more than one sentence; 33000 words across all transcripts).

This paper will discuss how our system implements the transfer of affect in a very limited range of metaphors. However, it should be noted that the system underlying our edrama ICA does not detect affect solely or even primarily via metaphor. Quite apart from the recognition of specifically emotive and affective lexis, the system deals with letter and punctuation repetition for emphasis ("yeessss," "!!!!"), interjections and onomatopoeia (grrrrrrr) (see Zhang et al. 2006 for details). However, these may be viewed as manifestations of an abstract conceptual metaphor that views or conceptualises 'more of some thing or some quality' as 'an increase along one salient dimension'; typically height. This often gives us the Lakovian conceptual metaphor MORE IS UP, but gives word length when dealing with text. The degree of increase is conveyed by our degree VNMA.

Our system uses a blackboard architecture, in which hypotheses arising from the processing go onto a central blackboard. The production of the various hypotheses can then be influenced by hypotheses posted by other processes, etc. In particular, we envisage metaphor processing being refined by using such information (see Smith et al. 2007 for more details).

2. Affect via Metaphor in an ICA

Our system currently detects and analyses the transference of affect in the cases where a human is cast as a non-human of various sorts, as in the following cases:

1) Casting someone as an animal. This often transfers some affect -negative or positive- from the animal to the human. Interestingly, since our attitude towards young or baby forms, regardless of the animal concerned, are typically affectionate, affection is often transferred, even when the adult form is negative ('pig: piglet', 'dog: puppy' etc.). We deal with animal words that have a conventional metaphorical sense but also with those that do not, for it may still be possible to note a particular affective connotation, and even if not, one can plausibly infer that some affect or other is being expressed without knowing if positive or negative.

2) Relatedly, casting someone as a monster, mythical creature or supernatural being of some sort, using words such as 'monster' itself, 'dragon,' 'angel,' 'devil.'

3) Relatedly, casting someone as an artefact, substance or natural object, as in 'Tom is a [sewer; real diamond; rock].

We currently do not deal with the related case of casting someone metaphorically as a special type of human, using words such as 'baby,' 'freak,' 'girl' [to a boy], 'lunatic'.

In addition, size adjectives (cf. Sharoff 2006) often convey affect. Thus, 'a little X' can convey affective qualities of X such as an affectionate attitude towards X, even if the X is usually negative as in 'little devils' to describe mischievous children (compare with the baby forms above), but may sometimes convey unimportance and contemptibility as in 'you little rat'. Similarly, 'big X' can convey the importance of X ('big event') or intensity of X-ness ('big bully') -and X can itself be metaphorical as in 'big baby' when said of an adult.

3. Metaphor Processing

The approach is split into two parts: recognition of potential metaphors; analysis of recognised elements to determine affect. Note that in some cases, e.g. using 'pig' as a negative term for a person, the metaphor analysis requires only lexical look-up (e.g., in WordNet, 2006). But, not all animal words have a person sense and as noted above baby forms often change the affect as do size

adjectives. Such cases motivate the further processing.

3.1 The Recognition Component

The basis here is a list of words/phrases (www.cs.bham.ac.uk/jab/ATT-Meta/metaphoricity-signals.html) we term 'metaphoricity signals', that often have metaphors as collocates. They include specific syntactic structures as well as lexical strings. We currently focus on three syntactic structures, 'X is/are a Y', 'You Y' and 'like [a] Y' and on the lexical strings, 'a bit of a', 'such a' and 'look[s] like'. Note that a distinction is often made between similes and metaphors, making the third structure a simile. Our view is that (many) similes represent just a particular way of expressing an underlying metaphorical connection between X and Y and so shouldn't be treated differently from the other realisations. In the user-testing transcripts, we judged signals as actually involving metaphor in the following proportions of cases: *X is/are a Y* – 38% (18 out of 47); *you Y* – 61% (22 out of 36); *a bit of a / such a* – 40% (but tiny sample: 2 out of 5). Also: *looks like* and *like* – 81% (35 out of 43). (Of course, metaphor is often not signalled and can occur in any syntactic form and not just the forms here.)

In order to detect signals, the Grammatical Relations (GR) output from the RASP parser (Briscoe et al. 2006) is used. This output shows typed word-pair dependencies between the words in the utterance. For example, the following three GRs are output for a sentence such as 'You are a pig', so allowing an 'X is a Y' signal to be detected.

```
|ncsubj| |be+_vbr| |you_ppy| |_|
```
(i.e. the subject of 'are' is 'you')

```
|xcomp| | be+_vbr| |pig_nn1|
```
(i.e. the complement of 'are' is 'pig')

```
|det| |pig_nn1| |a_at1|
```
(i.e. the determiner of 'pig' is 'a')

Note that the tags 'vbr' and 'ppy' are specific to 'are' and 'you', so we also detect tags for: 'is'; for 'he', 'she' and 'it'; and for proper and common nouns, as well.

The output for the 'You Y' structure is typically as in the following example:

```
|ncmod| |you_ppy| |idiot_nn1|
```

(with Y = 'idiot') making it possible to find the structure from that one relation. But a common problem with RASP on 'You Y' is that its 'Part of Speech' (POS) tagger seems to favour tagging Y as a verb, if it can. For example, the word 'cow' in place of 'idiot' is tagged as a verb. In such a case, our system looks the word up in the list of tagged words that forms part of the RASP tagger. If the verb can be tagged as a noun, the tag is changed, and the metaphoricity signal is detected. Once a syntactic structure resulting from metaphoricity signals is detected, the word(s) in Y position are pulled out to be analysed.

This approach has the advantage that whether or not the noun in the Y position has adjectival modifiers the GR between the verb and Y is the same so the detection tolerates a large amount of variation, an important desiderata for metaphor. Any such modifiers are found in modifying relations and can be extracted for later analysis.

For additional confidence we detect the lexical strings 'a bit of a' and 'such a'. 'Such a' is found using GRs of the following type:

```
|det| |idiot nn1| |an at1|
```
(i.e. the determiner of 'idiot' is 'an'.)

```
|det| |idiot nn1| |such da|
```
(i.e. the determiner of 'idiot' is 'such')

Note that 'idiot', is detected as a 'Y' type metaphor, independently of 'such a', by the syntactic structure detection process: the 'X-is-a-Y' metaphoricity signal. The 'a bit of a' strings are found similarly, but cause the complication that the word 'bit' is tagged as a noun, so will be pulled out as a metaphor word by the syntactic detection processes, instead of the intended Y word. If the 'a bit of a' string is then found, we pull out the noun relating to the 'of' that relates to 'bit', in this type of GR output:

```
|iobj| |bit_nn1| |of_io|

|dobj| |of_io| |idiot_nn1|
```

In addition to 'X is a Y' and 'You Y', another metaphoricity signalling syntactic structure is 'like Y'. This is found using GR's of the following type:

```
|dobj| |like_ii| |pig_nn1|
```

'like Y' is always found in this form, with the noun in question in the dobj (direct object) relation to 'like', and with an nn1 tag. This is inserted into the list of present metaphoricity signals, and an additional flag is raised if it is found in an 'X looks like Y' structure. The 'looks like' structure can be uncovered by spotting this GR:

```
|iobj| |look_vv0| |like_ii|
```

Detection of the 'looks like' structure is similar to 'such a' in that it is in addition to the main metaphoricity signal detection, in this case not only adding confidence, but also potentially altering the meaning and analysis of the metaphor.

The result of the recognition element is threefold: (1) a list of signals; (2) the X and Y nouns from the syntactic signals; (3) a list of words modifying that noun.

3.2 The Analysis Component

The analysis element of the processing that we shall discuss here takes the X noun (if any) and Y noun and uses WordNet 2.0 (2006) to analyse them. First, we try to determine whether X refers to a person (the only case the system currently deals with), partly by using a specified list of proper names of characters in the drama and partly by WordNet processing (The system also proceeds similarly if X is 'you'). If so, then the Y and remaining elements are analysed using WordNet's taxonomy. This allows us to see if the Y noun in one of its senses is a hyponym of (or member of the class of) animals, supernatural beings, substances, artefacts or natural objects. If this is established, the system sees if another of the senses of the word is a hyponym of the person synset, as many metaphors are already given as senses in WordNet. If the given word contains different synsets or senses that are hyponyms of both animal etc. and person, then we search for evaluative content about the metaphor.

We have developed a method of automatically detecting the evaluation of a given metaphorical sense of a word. Intermediate synsets between the metaphorical sense of the given word and the person synsets contain glosses, which are descriptions of the semantic content of a synset. For example, the gloss of the synset of 'shark' that is a hyponym of 'person' is "a person who is ruthless and greedy and dishonest"; that of 'fox' is "a shifty deceptive person". We search the words and glosses from the intermediate synsets for words that indicate a particular affective evaluation. This is somewhat crude, since we do not parse the glosses, although a limited parser is currently being implemented. Consequently, both 'evil' and 'not evil' if found in a gloss will be taken to indicate a negative evaluation. (see Veale 2003 for related use of WordNet glosses).

Now there exist numerous lists and resources containing evaluative words. Indeed, SentiWordNet (Esuli & Sebastiani, 2006) is based on the glosses of the WordNet synsets and assigns three numerical scores describing how objective, positive, and negative the terms contained in the synset are. See also WordNet-Affect (Strapparava et al. 2004) However, in practice we found that very many of the animals etc. we wished to assign a positive or negative evaluation to were given a neutral score in SentiWordNet and so we created our own list. We decided that since we were searching though WordNet glosses, it would be most appropriate to create a list from WordNet itself. This we did in the following manner. WordNet contains a 'quality' synset which has 'attribute' links to four other synsets, 'good', 'bad', 'positive' and 'negative'. We are currently only looking for positive or negative affective evaluations, so this group of synsets provides a core set of affect indicating words to search for in the intermediate nodes. This set is expanded by following WordNet's 'see also' links to related words, to produce lists of positivity and negativity indicators. For example, 'bad' has 'see also' links to five synsets, including 'disobedient' and 'evil'; we then look up the 'see also' links in these five synsets and include these related words in the 'bad' list, and so on, through five iterations, producing a list of over 100 words indicating negativity.

With this list, we can search through the words and glosses from the intermediate nodes between the given metaphor synset (arising from the Y component in the sentence) and 'person', tallying the positivity and negativity indicating words found. We can then assign the affective evaluation of the metaphor, so having more negativity indicators than positivity indicators suggests that when the word is used in a metaphor it will be negative about the target. If the numbers of positivity and negativity indicators are equal, then the metaphor is labelled positive or negative, implying that it has an affective quality but we cannot establish what. This label is also used in those examples where an animal does not have a metaphorical sense in WordNet as a kind of person (for example, 'You elephant' or 'You toad').

It might be thought that the need for an additional person hypernym for Y is not necessary and that a search through the glosses of just the animal etc synsets in the hypernym tree Y would yield a relevant affective evaluation, at least in cases where there is no additional person sense. But this appears not to be the case. The glosses tend to be technical with few if any affective connotations. For example, 'toad' surprisingly does not have an alternative person sense in WordNet. The glosses of its 'amphibian, vertebrate and chordate hypernyms give technical information about habitat, breeding, skeletal structure, etc. but nothing affective. Worse still, false friends can be found. Thus, the word 'important' is used in many glosses in phrases like 'important place in the food chain' and this consequently causes some strange positive evaluations (for example of 'Cyclops' or 'water fleas').

We noted earlier that baby animal names can often be used to give a statement a more affectionate quality. Some baby animal names such as 'piglet' do not have a metaphorical sense in WordNet. In these cases, we check the word's gloss to see if it is a young animal and what kind of animal it is (The gloss for piglet, for example, is "a young pig").We then process the adult animal name to seek a metaphorical meaning but add the quality of affection to the result. A higher degree of confidence is attached to the quality of affection than is attached to the positive/negative result, if any, obtained from the adult name. Other baby animal names such as 'lamb' do have a metaphorical sense in WordNet independently of the adult animal, and are therefore evaluated as above. They are also tagged as potentially expressing affection, but with a lesser degree of confidence than that gained from the metaphorical processing of the word. However, the youth of an animal is not always encoded in a single word: e.g., 'cub' may be accompanied by specification of an animal type, as in 'wolf cub'. An extension to our processing would be required to handle this and also cases like

'young wolf' or 'baby wolf'.

If any adjectival modifiers of the Y noun were recognized the analyser goes on to evaluate their contribution to the metaphor's affect. If the analyser finds that 'big' is a modifying adjective of the noun it has analysed, the metaphor is marked as being more emphatic. If 'little' is found the following is done. If the metaphor has been tagged as negative and no degree of affection has been added (from a baby animal name, currently) then 'little' is taken to be expressing contempt. If the metaphor has been tagged as positive OR a degree of affection has been added then 'little' is taken to be expressing affection. These additional labels of affection and contempt are used to imply extra positivity and negativity respectively.

4. Examples of the Course of Processing

In this section we discuss three examples in detail and seven more with brief notes.

4.1 You piglet

1). The metaphor detector recognises the 'You Y' signal and puts the noun 'piglet' on the blackboard.
2). The metaphor analyser reads 'piglet' from the blackboard and detects that it is a hyponym of 'animal'.
3). 'Piglet' is not encoded with a specific metaphorical meaning ('person' is not a hypernym). So the analyser retrieves the gloss from WordNet.
4). It finds 'young' in the gloss and retrieves all of the words that follow it. In this example the gloss is 'a young pig' so 'pig' is the only following word. If more than one word had followed, then the analysis process is repeated for each of the words following 'young' until an animal word is found
5). The words and glosses of the intermediate nodes between 'pig' and 'person' contain 0 positivity indicating words and 5 negativity indicating words, so the metaphor is labelled with negative polarity.
6). This example would result in the metaphor being labelled as an animal metaphor which is negative but affectionate with the affection label having a higher numerical confidence weighting than the negative label.

4.2 Lisa is an angel

1). The metaphor detector recognises the 'X is a Y' signal and puts the noun 'angel' on the blackboard. 'Lisa' is recognised as a person through a list of names provided with the individual scenarios in e-drama.
2). The metaphor analyser finds angel that it is a hyponym of 'supernatural being'.
3). It finds that in another of its senses the word is a hyponym of 'person'.
4). The words and glosses of the intermediate nodes between 'angel' and 'person' contain 8 positivity indicating words and 0 negativity indicating words, so the metaphor is labelled with positive polarity.
5). This example results in the metaphor being labelled as a positive supernatural being.

4.3 Mayid is a rock

1). The metaphor detector recognises the 'X is a Y' signal and puts the noun 'rock' on the blackboard. 'Mayid' is recognised as a person through a list of names provided with the individual scenarios in e-drama.
2). The metaphor analyser finds rock is a hyponym of 'natural object'.
3) It finds that in another of its senses the word is a hyponym of 'person'.
4). The words and glosses of the intermediate nodes between 'rock' and 'person' contain 4 positivity indicating words and 1 negativity indicating words, so the metaphor is labelled with positive polarity.
5). This example would result in the metaphor being labelled as a positive natural object.

4.4 Other Examples

1). 'You cow': this is processed as a negative animal metaphor. The synset of 'cow' that is a hyponym of 'person' has the gloss "a large unpleasant woman". Interestingly, 'large' is included in the list of positivity indicators by the current compilation method, but the negativity of the metaphor is confirmed by analysis of the intermediate synsets between 'cow' and 'person', which are 'unpleasant woman', 'unpleasant person' and 'unwelcome person'. These synsets, along with their glosses, contain six negativity and just one positivity indicator.
2). 'You little rat': this animal metaphor is determined as negative, having three senses that are hyponyms of 'person', containing three positivity indicators and five negativity indicators. 'Little' provides an added degree of contempt.
3). 'You little piggy': 'piggy' is recognized as a baby animal term and labelled as expressing affection. The evaluation of 'pig' adds a negative label, with no positivity indicators and three negativity indicators, and 'little' adds further affection since the metaphor already has this label from the baby animal recognition. This is therefore recognized as a negative metaphor but meant affectionately.
4). 'You're a lamb': recognized as an animal metaphor and a young animal. It has an 'affectionate' label and is recognized as a positive metaphor, with its two senses that are hyponyms of 'person' contributing two positivity indicators and one negativity indicator. The negative word in this case is 'evil', coming from the gloss of one of the intermediate synsets, 'innocent': "a person who lacks knowledge of evil". This example highlights a failing of using individual words as indicators: negations within sentences are not recognized.
5). 'You are a monster': one sense of monster in WordNet is a hyponym of animal. Therefore, this is recognized as an animal metaphor, but affect evaluation reveals three negativity and three positivity indicators, so it is analysed as 'positive or negative'. These indicators are found in two opposed senses of monster: 'monster, fiend, ogre': "a cruel wicked and inhuman person" (analysed as negative); and 'giant, monster, colossus': "someone that is

abnormally large and powerful" (analysed as positive, due to 'large' and 'powerful').
6). 'She's a total angel': a positive supernatural being metaphor, with eight positivity indicators and no negativity indicators from two senses that are hyponyms of 'person', but currently 'total' makes no contribution.
7). 'She is such a big fat cow': a negative animal metaphor made more intense by the presence of big. It has an extra level of confidence attached to its detection as two metaphoricity signals are present but currently 'fat' makes no contribution.

5. Conclusions and Further Work

The paper has discussed a relatively 'shallow' type of metaphor processing, although our use of robust parsing and complex processing of a thesaurus take it well beyond simple keyword approaches or bag-of-words approaches. Note that we do not wish simply to 'precompile' information about animal metaphor (etc.) by building a complete list of animals (etc.) in any particular version of WordNet (and also adding the effects of potential modifiers such as 'big' and 'little'), because we wish to allow the work to be extend to new versions of WordNet and to generalize as appropriate to thesauri other than WordNet, and because we wish to allow ultimately for more complex modification of the Y nouns, in particular by going beyond the adjectives 'big' and 'little'. We recognize that the current counting of positive and negative indicators picked up from glosses is an over-simple approach, and that the nature of the indicators should ideally be examined. This is a matter of both ongoing and future research. The processing capabilities described make particular but nonetheless valuable and wide-ranging contributions to affect-detection for ICAs. Although designed for an edrama system, the techniques plausibly have wider applicability. The development of the processing in a real-life application is also enriching our basic research on metaphor, such as the role of VNMAs.

6. Acknowledgements

This work has been supported by EPSRC grant EP/C538943/1 and grant RES-328-25-0009 from the ESRC under the ESRC/EPSRC/DTI "PACCIT" programme. We wish to thank our colleagues Catherine Smith and Sheila Glasbey.

7. References

Agerri, R., Barnden, J., Lee, M., Wallington, A. (2007). Metaphor, inference and domain independent mappings In *International Conference on Recent Advances in Natural Language Processing (RANLP 2007,* G Angelova, K Bontcheva, R Mitkov, N Nicolov, (Eds), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 07), Borovets, Bulgaria, pp. 17-24.

Barnden, J., Glasbey, S., Lee, M., Wallington, A. (2004) Varieties and directions of interdomain influence in metaphor, *Metaphor and Symbol*, 19, pp.1-30.

Briscoe, E., Carroll, J. and Watson, R. (2006) The Second Release of the RASP System. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Sydney.

Esuli, A & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining, In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006),* Genova, Italy.

Fass, D. (1997). *Processing Metaphor and Metonymy*. Greenwich, Connecticut: Ablex.

Fussell, S., & Moss, M. (1998). Figurative Language. In S.R. Fussell and R.J. Kreuz (Eds) *Emotional Communication. Social and Cognitive Approaches to Interpersonal Communication*. Hillside, NJ: Lawrence Erlbaum, pp. 113-142.

Gibbs, R. (1992) Categorization and metaphor understanding. *Psychological Review*. 99(3) 572-577

Hobbs, J. (1990). *Literature and Cognition*. Center for the Study of Language and Information, Stanford University: CSLI Lecture Notes, 21.

Martin, J. (1990). *A Computational Model of Metaphor Interpretation*. San Diego, CA: Academic Press.

Mason, Z. (2004). CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1).

Kövecses, Z. (2000). *Metaphor and Emotion: Language, Culture and Body in Human Feeling*. Cambridge: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of the National Conference on Artificial Intigence*. AAAI Press.

Picard, R.W. (2000). *Affective Computing*. Cambridge MA: The MIT Press.

Reddy, M.J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A Ortony, (Ed) 2nd edn 1993 *Metaphor and Thought*. Cambridge: Cambridge University Press, 164-201.

Sharoff, S. (2006). How to Handle Lexical Semantics in SFL: a Corpus Study of Purposes for Using Size Adjectives. In S Hunston & G Thompson (Eds) *Systemic Linguistics and Corpus*. London: Equinox, pp 184-205.

Smith, C.J., Rumbell T.H., Barnden, J.A., Lee, M.G., Glasbey, S.R., & Wallington, A.M. (2007). Affect and Metaphor in an ICA: Further Developments. In Pelachaud, C.; Martin, J.-C.; André, E.; Chollet, G.; Karpouzis, K.; Pelé, D. (Eds.) *Intelligent Virtual Agents. 7th International Working Conference, IVA 2007*. Lecture Notes in Computer Science, Vol.4722. Heidelberg:,Springer-Verlag. pp 405-406

Strapparava, C. & Valitutti, V. (2004). WordNet-Affect: An Affective Extension of WordNet, In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* Lisbon, Portugal. 1083-1086.

Veale, T. (2003). Systematicity and the Lexicon in Creative Metaphor. In *Proceedings of the ACL Workshop on Figurative Language and the Lexicon, the 41st Annual Association for Computational Linguistics Conference* (ACL 2003), Sapporo, Japan.

Wallington, A, Barnden, J., Glasbey, S., Lee, M. (2006). Metaphorical reasoning with an economical set of mappings. *Delta* 22 (especial). pp 147-171

WordNet (2006). *WordNet, A Lexical Database for the English Language. Version 2.1* Cognitive Science Laboratory. Princeton University.

Zhang, L., Barnden, J.A., Hendley, R.J. & Wallington, A.M. (2006). Exploitation in Affect Detection in Improvisational E-drama. In J Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds) *Proceedings of the 6th International Conference on Intelligent Virtual Agents*. Lecture Notes in Computer Science, 4133, Springer, pp.68-79.

The 'return' and 'volatility' of sentiments: An attempt to quantify the behaviour of the markets?

Khurshid Ahmad
Computer Science
Trinity College Dublin
khurshid.ahmad@cs.tcd.ie

Abstract

A study of changes in 'sentiment' about the booming Celtic Tiger, a term associated with the economic performance of the Irish Republic during the 1990's and in the first half of this decade, was conducted by analysing a corpus of news paper stories published in an authoratitive and established Irish newspaper (*The Irish Times*) during (1995-2000) and supposedly after the boom (2001-2005). The corpus was designed to only incorporate texts in which one or more affect word, in an arbitrary collection of affect words, co-occurs with three terms: 'Irish' or 'Ireland', and *economy* in the same news story. I present a method for quantifying changes using (the logarithm) of the ratio of the frequency of positive affect words (and for negative words) for two successive months: from these 'returns', a term used in financial economics and corporate finance, we computed the so-called 'volatility' in the two time series – the standard deviation of the positive and negative return time-series. 'Return' and 'volatility' computations are typically carried out on the prices of financial instruments (e.g. shares and commodities) for estimating the risk associated with the instruments. We show that there is some agreement between periods of high-volatility in a stock market, or a selected stock market index to be precise, and the volatility of 'bad' news during the year. I have used a corpus of texts, 2.6 million words in total, retrieved from *The Irish Times* Digital Archive and published between 1995-2005 and used the Harvard Dictionary of Affect to create the arbitrary collection of positive and negative affect words. The use of the returns and volatility for word frequency changes, if found to be methodologically sound and correlative to changes in prices or indices, can then be used to compute the *risk* associated with a financial market in the well-grounded use of the two metrices in finance studies and in international economics.

1. Introduction

Sentiment analysis is now becoming an established tool for the analysis of financial and commodity markets. The roots of this subject lie in the earlier work (c. 1950's) on content analysis on the one hand and on the other in work on bounded rationality and 'herd behaviour' by Herbert Simon and Daniel Kahnemann. Information extraction and corpus linguistics have been used in extracting the distribution of the so-called affect words and their collocates.

We begin this paper with a mini (or nano) tutorial on the two key metaphorical terms used in finance studies, especially in the study of changes of prices and that of the value of the indices of a market: *return* and *volatility*. We briefly describe some related work in sentiment analysis. This is followed by a description of a corpus-based study of the variation in the frequency of positive and negative words, as defined in the Harvard *Dictionary of Affect*. An afterword concludes the paper.

2. Metaphors of 'Return' and of 'Volatility'

The literature on financial economics that is closely related to the analysis of market sentiment frequently refers to two key terms: *return* and *volatility*. These terms have retained much of their original meaning that is, 'return' broadly refers to the 'act of coming back', as established upon the entry of this word in the English language in the 14th century or thereabouts. This word has been adapted, or to put loosely has been used as a metaphor for coming back as in the definition: 'Pecuniary value resulting to one from the exercise of some trade or occupation; gain, profit, or income, in relation to the means by which it is produced'. In financial economics a return, or 'price change quantity' is defined as the logarithm of the ratio of the current and past price (or an index of a stock exchange or any other aggregated index,

like Standards & Poor, Financial Times-Stock Exchange Index):

Let p_t be the price today and p_{t-1} the price yesterday, so the return r_t is defined as:

$$r_t = \log(p_t / p_{t-1})$$

The word 'volatility' is much more graphic as it started its journey into the English language from Latin in the 17th century: 'The quality, state, or condition of being volatile, in various senses' and the metaphorical use of this word includes references to 'tendency to lightness, levity, or flightiness; lack of steadiness or seriousness'. Benoit Mandelbrot (1963) has argued that the rapid rate of change in prices (the *flightiness* in the change) can and should be studied and not eliminated – 'large changes [in prices] tend to be followed by large changes –of either sign- and small changes tend to be followed by small changes'. The term *volatility clustering* is attributed to such clustered changes in prices. Mandelbrot's paper drew upon the behaviour of commodity prices (cotton, wool and so on), and 'volatility clustering' is now used for almost the whole range of financial instruments (see Taylor 2007 for an excellent and statistically well-grounded, yet readable, account of this subject).

There are different kinds of measures of volatility, a commonly used version is called *realized* or *historical volatility.* Volatility (ν) of a stock price or the value of an index is defined over a trading period *n* and is the standard deviation of the past returns ($\log(p_t / p_{t-1})$):

$$\nu = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(r_{t-k} - \bar{r})^2}$$

where $\bar{r}$ is the average value of past returns in the period *n.* Econometricians have observed that it is perhaps easier to study to conduct a statistical analysis of changes in market prices as there is little or no correlation between consecutive price changes; there is a significant correlation in the prices. It has been argued by Robert Engle, the 1993 co-winner of the Nobel prize in economics, that '[a]s time goes by, we get more information on these future events and re-value the asset. So at a basic level, financial price volatility is due to the arrival of new information. Volatility clustering is simply clustering of information arrivals. The fact that this is common to so many assets is simply a statement that news is typically clustered in time.' (1993:330).

The term *news arrivals* or *information arrivals* is defined rather differently in the literature in finance. The standard practice in financial economics is either to use daily counts of news stories as a proxy for information arrival or, in simulations, a random number generator is used for generating the number of news arrivals (Bauwens, Omrane and Giot 2005). This method has been refined to count only those stories that comprise a given (set of) keyword(s) (DeGennaro and Shrieves 1997, Chang and Talyor 2003), and more recently Tetlock (2007) has used affect words in Stone et al. (1996) *General Inquirer* Lexicon to correlate market movements with change in the frequency of affect words. More of this later.

Let us look at the concepts of return and volatility in the context of *information arrivals* by looking at the recurrence of news items containing the same keywords and, more importantly, on the recurrence of the same metaphorical terms. Consider the use of the term *credit crunch*: According to Wikipedia (2008) the term is defined as 'a sudden reduction in the availability of loans (or "credit") or a sudden increase in the cost of obtaining a loan from the banks.'. This term has been around for over 30 years or more, but let us look at the usage of this term since 1981: The *New York Times* archives were searched for the compound term "credit crunch" and over 400 articles, published between 1981-2008, were found that contained the term. There may have been other articles which the NYT authorities did not or could not include, but the NYT is considered as an authoritative and usually un-biased source of US and international political and economic reporting. The number of stories containing the term appear to have a 5 year cycle, except in the last decade where "credit crunch" only appeared in large number in 2007; between January and April 2008 there were 54 stories compared to the whole of 2007 where there were 77 stories in all. My projected value of the number of stories in 2008 is 156, based on the current average of about 13 a month (Table 1)

Table 1 The number of stories per year comprising the term *credit crunch* that appeared in (or are in the archive of) *New York Times*

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	Decade Total
# Stories	**20**	7	4	4	**8**	5	3	3	6	**59**	119
Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	
# Stories	**84**	37	14	4	**11**	3	2	**47**	3	**3**	208
Year	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**			
# Stories	**10**	4	1	1	**1**	1	**77**	**54** (156)			149

Let us look at the 'returns' of the number of stories in Table 1 and compute the annual historical volatility purely on the basis of the changes in the numbers of stories in *New York Times*. It has been argued that volatility, computed using return of prices or index values, increases during crises periods, say during the 1929 US Great Depression, the period leading up to the resignation of the US President Nixon in 1974, and the days after the 2001 9/11 attacks (see Taylor 2005:191-93).

We have plotted the number of stories per year, the consecutive year returns, and the volatility for a reporting period of 5 years. There is a much greater variation in the returns (based on the current and past numbers of news stories containing the term 'credit crunch') when compared to the fluctuations recorded in the actual number of stories. One quantification of such a fluctuation is volatility or the standard deviation of past returns. The volatility increased every 5 years until 2000 – indicating an *increasing* sense of 'crises'. Volatility decreases during calmer periods. In our extremely simple illustration, we note that the volatility decreased during 2001-2005 – a period of massive growth partly fuelled by 'easy credit'- and lastly for the three years (2006-2008) the volatility has increased dramatically incorporating the period of the very frequent use of the term 'credit crunch'. (Figure 1):

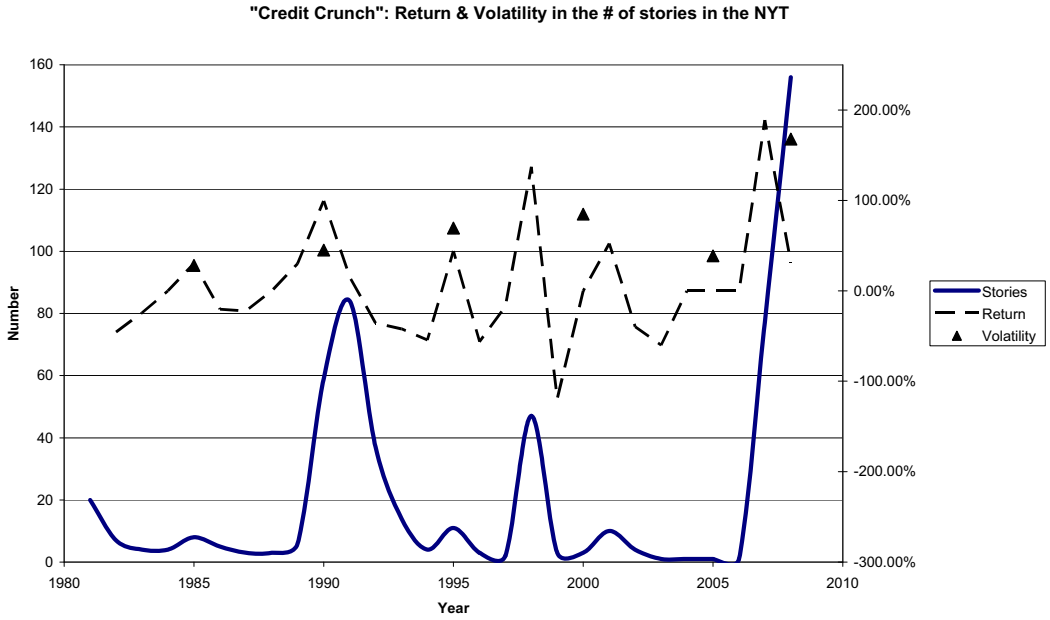

Figure 1. The number of stories are on the left vertical axis, and the percentage change in returns and volatility are on the right. Note that I have used the extrapolated number of stories for the whole of 2008, 13 per month is the norm which has been extrapolated to 156 for the year.

The above figure shows that a single statistic, volatility, can be used in the quantification of rapid changes. The questions is this: will this statistic throw any light on the changes in market sentiment, based on methods in sentiment analysis that use the frequency of positive and negative affect words as a measure of sentiment?

3. Sentiment Analysis

3.1. The roots of computational sentiment analysis

Sentiment is defined as "an opinion or view as to what is right or agreeable" and political scientists and economists have used this word as a technical term. When sentiments are expressed through the faculty of language, we tend to use certain literal and metaphorical words to convey what we believe to be right or agreeable. There are a number of learned papers and reviews in computational sentiment analysis that are available (Kennedy and Inkpen 2006 and references therein)

One of the pioneers of political theory and communications in the early 20th century, Harold Lasswell (1948), has used sentiment to convey the idea of an attitude permeated by feeling rather than the undirected feeling itself. (Adam Smith's original text on economics was entitled *A Theory of Moral Sentiments.*) Namenwirth and Laswell (1970) looked at the Republican and Democratic party platforms in two periods 1844-64 and 1944-64 to see how the parties were converging and how language was used to express the change. Laswell created a dictionary of affect words (*hope, fear*, and so on) and used the frequency counts of these and other words to quantify the convergence.

This approach to analysing contents of political and economic documents – called content analysis- was given considerable fillip in the 1950's and 1960's by Philip Stone of Harvard University who created the so-called General Inquirer System (Stone et al 1996 and Kelly and Stone 1975) and a large digitised dictionary – the Harvard *Dictionary of Affect* comprising over 8,500 words carefully selected using a criterion developed by the psychologist Charles Osgood including positive/negative words, words to express strength and weakness, and words to describe activity and arousal (Stone's dictionary

includes a number of entries used by Harold Laswell; these entries are thus labelled).

Recently, the digitised Harvard *Dictionary of Affect* has been used to 'measure' sentiment in the financial markets. Tetlock (2007) has analysed a commentary column in the *Wall Street Journal* using the Harvard *Dictionary of Affect* and correlated the frequency of affect words with trading volumes of shares in the New York Stock Exchange: He concludes that 'high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume.' This is amongst the first reported studies in financial sentiment analysis that is rooted strongly in econometric analysis (especially through the use of auto-regressive models in the framework of conditional heteroskedacity) that has analysed the contents of the news in conjunction with the study of information arrival (see also Lidén 2006). Tetlock's selection of comment or opinion in a newspaper, classified as imaginative writing rather than the informative news reportage, may raise some methodological questions in text analysis about whether or not opinions alone comprise a representative sample of texts that has been used for analysis (see, for example Althaus, Edy and Phalen 2001, Ahmad 2007b)

3.2. A corpus-based study of sentiments, terminology and ontology over time

We report on some work recently carried out on compiling a representative sample of texts, in a given domain, that can be used for analysing sentiments. Once the corpus is compiled we then extract terminology that is used in the domain automatically and statistically significant collocates of the candidate domain terminology are used in the construction of a candidate ontology (see Ahmad 2007a for details). In my previous work I have avoided using pre-compiled dictionaries of affect and used the so-called 'local grammar' constructs for extracting patterns that were 'sentiment-laden' (Almas and Ahmad 2006) – an approach that has allowed us to look for sentiment in texts in typologically diverse languages like English (and Urdu), Chinese and Arabic. For the purposes of comparison with other work in financial sentiment analysis, I have used the Harvard *Dictionary of Affect*. Perhaps, the innovative aspect of this paper is the fact that I have computed returns and volatility of affect in a corpus drawn for a representative newspaper website. My hypothesis is this: Can the computation of the volatility of affect, found in news paper reportage and editorials, help in quantifying (financial and economic) risk, much in the same manner risk computations based on prices and values of index help in quantifying risk?

3.3. Corpus Preparation and Composition

The design of the corpus was motivated by the state of Irish economy during the period of 1995-2005; first five years were the so-called Celtic Tiger boom eras (1996-2000) and the next five year period comprised the dot.com bubble, September 2001 attacks and the consequent down turn and the lead upto the introduction of the Euro. The authoritative and influential *Irish Times* that has been published since the 1850's and has a digital archive going back to 1859. One of my student (Nicholas Daly) used a text-retrieval robot to search and retrieve all items (news reports, editorials, or op-ed columns) based on the robot user: We chose the period (1995-2005) and gave the robot three keywords: *Ireland/Irish* and *Economy.* The corpus comprises 2.6 million words distributed over 4075 news reportage and editorial items (Table 2):

Table 2 Distribution of stories in our *Irish Times Corpus*

Year	No. of Stories	No. of Words	Year	No. of Stories	No. of Words
1996	296	165937	2001	562	360026
1997	395	259748	2002	367	256613
1998	465	296531	2003	377	250415
1999	447	295873	2004	377	250376
2000	462	306063	2005	327	234101
TOTAL	2065	1324152		2010	1351531

The size of the year, viewed on an annual basis, appears to be comparable (Mean=407,Standard Deviation=51522): only in two years both the number of stories and the verbiage was above one-standard deviation above the mean (1996 and 2001), and the number of stories in 2005 were just one s.d. above the mean (1.04).

3.4. Candidate Terminology and Ontology

We found that 'sentiment' in itself was a keyword and analysis of its statistically significant collocates showed that despite the boom in the late 90's the focus of *Irish Times* content was on more negative aspects, but the next 5 years show the establishment of a whole terminology nucleating around 'sentiment' (Table 3):

Table 3: Compound words with 'sentiment' as a head word – a comparison over 5 year periods:

1995-2000	2001-2005
sentiment investor_sentiment factors_affecting_investor_sentiment market_sentiment bond_market_sentiment negative_sentiment poor_sentiment	sentiment business_sentiment consumer_sentiment consumer_sentiment_survey irish_consumer_sentiment investor_sentiment sentiment_index sentiment_index_climbed sentiment_index_produced sentiment_index_rose sentiment_index_surpassed

The above analysis was carried out using the computation of significant collocates following Frank Smadja (1993) and the assumption here was that if the word sentiment is to the left of a another word, excluding the so-called closed class words, then sentiment is the headword. The output was processed using the ontology system Protégé.

3.5. Historical Volatility in our Corpus

The 2.6 million word corpus was analysed by computing the frequency of affect words in Harvard *Dictionary of Affect* (H-DoA) that were present in the texts in the corpus. The frequency was normalised for the length of the individual texts. The H-DoA comprises a large number of categories as mentioned above: we have used only two categories *Positive* and *Negative* affect word categories that respectively has 1916 and 2292 words. For each news item on a given day, the frequency of all words that were labelled *Positive and Negative* in the H-DoA was computed. The frequency counts were aggregated on a monthly basis and returns computed. The standard deviation of the returns on annual basis was calculated and we then had *volatility* of 'positive' sentiments and that of the 'negative' sentiments.

The first thing to notice about our results that the 'return' (change in frequency) shows much greater fluctuation in value than the frequency itself; this confirms the findings in econometrics in the context of prices and the change in prices (see, for example, 2003). This is true of both the negative and positive word frequency time-series, despite the preponderance of positive words over the negatives. (Figures 2a and 2b)

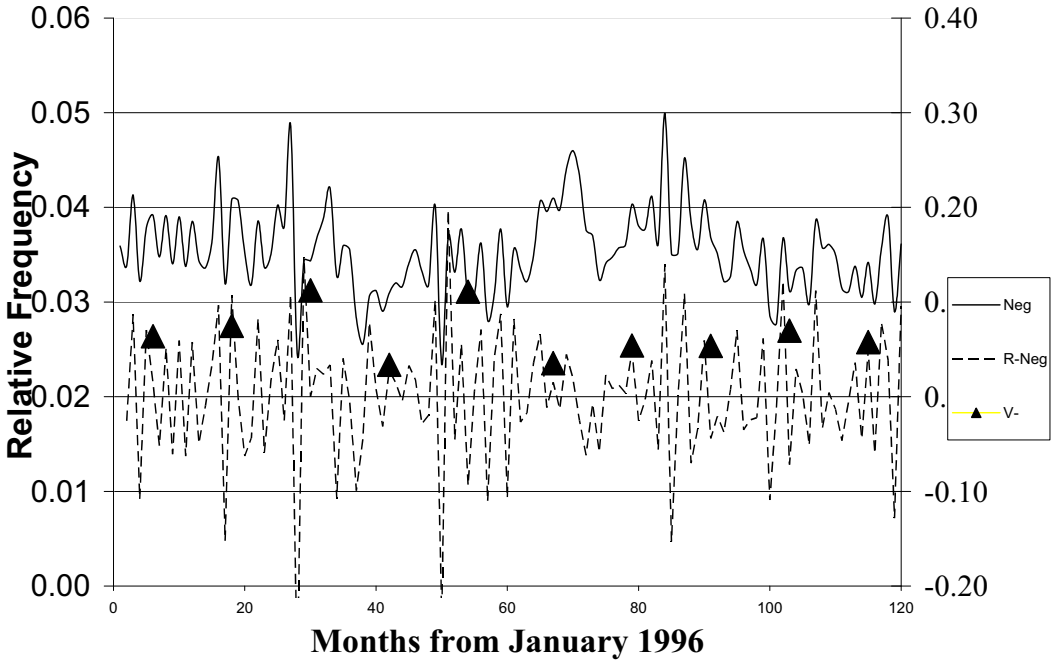

Figure 2a. Changes in the frequency (full line) of negative affect terms in our *Irish Times Corpus* (displayed monthly for 1996-2006). The returns are shown in dashed line (and values on the vertical axis on the right hand. The historical volatility is indicated by solid triangles and values are on the left-hand vertical axis.

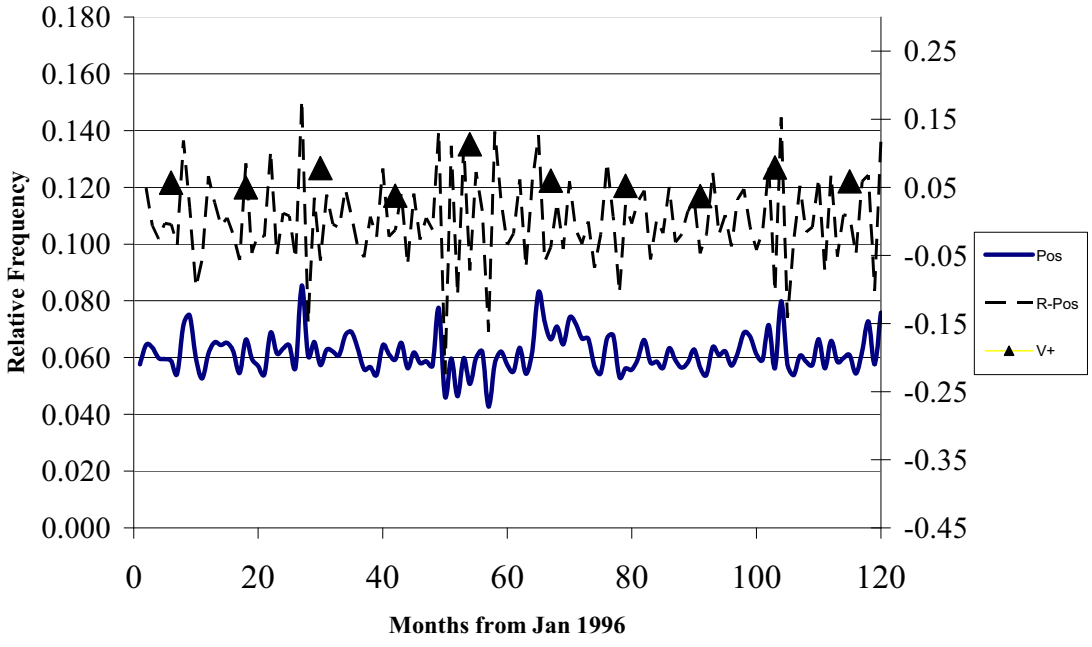

Figure 2b. Changes in the frequency (full line) of negative affect terms in our *Irish Times Corpus* (displayed monthly for 1996-2006). The returns are shown in dashed line (and values on the vertical axis on the right hand. The historical volatility is indicated by solid triangles and values are on the left-hand vertical axis.

The changes in historical volatility computed over the ten year period shows interesting results: in 1998 and 2000 the negative series had a higher than 'normal' volatility (one standard deviation above the norm) and in the two intervening years the volatility was below the norm (1999 and 2001). The positive affect series has below the norm volatility in 1999 and 2003 and much higher volatility in 2000 (2 standard deviations) (see Table 4).

Table 4. Volatility changes in our two time series

Year	Volatility		Volatility	
	Negative	Std Dev	Positive	Std Dev
1996	0.064		0.057	
1997	0.075		0.050	
1998	0.112	1.7	0.078	
1999	0.034	-1.2	0.038	-1.1
2000	0.111	1.6	0.113	2.2
2001	0.036	-1.1	0.060	
2002	0.054		0.052	
2003	0.054		0.037	-1.1
2004	0.070		0.079	
2005	0.058		0.059	

Finally, we show the variation in the Irish Stock Exchange Index of 100 top companies listed on the Exchange (ISEQ 100). We have had access to the values of the Index on a daily basis and we have used the value at the end of the month of each year as the ISEQ Index value and then computed returns and volatility for the period 1996-2005.

The volatility in ISEQ is smaller in comparison with that in the negative and positive affect series: this may be an artefact of computation as is the considerable variation in the volatility series of affect when compared to ISEQ (see Figure 5).

4. Afterword

The above results give us some sense of how to find sentiment words and quantify the changes in sentiment. It is perhaps too early to read the runes: whether we can use the volatility of affect times series to compute risk? I am yet to confirm or reject my hypothesis of the possible use of sentiment volatility in risk computations. But the study looks promising. We are looking at the auto-correlation in the various time series and computing other econometric metrices to quantify changes in sentiment.

In a related study, myself and my colleagues are looking at the effect of the use of different dictionaries of affect on the measurement of sentiments, including that of the H-DoA. We hope to use the system for analysing reports about emerging markets and specific financial instruments (shares, derivatives, bonds) and commodities: we intend to go beyond the *professional media* (newspapers, company documents, stock exchange reports) and include *social media* (blogs, e-mails and contrarian reports). It is through the social media that the contagion affecting the stock markets spreads. This project is undertaken jointly with Trinity Business School and the Irish Stock Exchange.

A sentiment analysis based on the indirect evidence of social and professional media is only one part of the overall picture. The Trinity Sentiment Analysis Group, a multi-disciplinary group including computer scientists, linguists and economists, has launched a sentiment survey for Irish institutional and individual investors. This survey was originally developed by Robert Shiller of Yale University International Centre of Finance; we have launched this Survey in collaboration with Yale[1]. The work of the Trinity Sentiment Group is ambitious and is focussed on engendering an openness and transparency in the workings of the vitally important financial sector. We are endeavouring to bring together and synthesise inputs from the professional media, the social media, data from the stock markets, and views of the stakeholders in a common framework. This is a long term program of work which we have just begun.

Acknowledgments

Dr Ann Devitt worked with me on this project and she has been looking at different dictionaries of affect. She had written the program that

[1] https://www.cs.tcd.ie/Khurshid.Ahmad/SurveySite/

computes affect word frequency. Dr Chaoxin Zheng and Daniel Iseman have helped with the layout and proof-reading.

Figure 5. Changes in the historical volatility in the affect series and in the ISEQ Index

References

Ahmad, Khurshid (2007a). 'Artificial Ontologies and Real Thoughts: Populating the Semantic Web? (Invited Keynote Speech). In (Eds). Roberto Basili & Maria Teresa Pazienza *Lecture Notes on Artificial Intelligence Series (4733) – Proc. 10th Ann. Congress of Italian Ass. of Art. Intelligence.* Berlin: Springer-Verlag. pp 3-23.

Ahmad, Khurshid (2007b). 'Being in Text and Text in Being: Notes on Representative Texts'. In (eds.) Gunilla Anderman and Margaret Rogers. Incorporating Corpora: *The Linguist and the Translator, Clevedon, England:Multilingual Matters.* pp 60 - 94

Almas, Yousif., and Khurshid Ahmad. (2007). A note on extracting 'sentiments' in financial news in English, Arabic & Urdu, *The Second Workshop on Computation, al Approaches to Arabic Script-based Languages*, Linguistic Society of America 2007 Linguistic Institute, Stanford University, Stanford, California., July 21-22, 2007, Linguistic Society of America. pp 1 - 12

Althaus, Scott, Jill Edy, and Patricia Phalen. 2001. "Using Substitutes for Full-Text News Stories in Content Analysis: Which Text is Best?" *American Journal of Political Science* 45(3): 707-723.

Bauwens, L.; W.B. Omrane; and P. Giot. (2005). "News Announcements, Market Activity and Volatility in the Euro/Dollar Foreign Exchange Market." *Journal of International Money and Finance*. Vol. 24, pp 1108-1125.

Chang, Y., and Taylor, S.J. (2003) Information Arrivals and Intraday Exchange Rate Volatility. *Journal of International Financial Markets, Institutions and Money*, vol 13, pp85-112

DeGennaro, R., and R. Shrieves (1997): "Public information releases, private information arrival and volatility in the foreign exchange market" . *Journal of Empirical Finance* Vol 4, pp 295–315.

Engle III, Robert. (2003). *Econometric models and financial practice.* http://nobelprize.org/nobel_prizes/economics/laureates/2003/engle-lecture.html

http://en.wikipedia.org/wiki/Credit_crunch (Accessed 17th April 2008)

Kelly, Edward., and Stone, Philip. (1975). *Computer Recognition of English Word Senses. Amsterdam*: North-Holland Linguistic

Series
Kennedy, Alistair & Inkpen, Diana. (2006). Sentiment classification using contextual valence shifters. *Computational Intelligence*. Vol. 22 (No. 2), pp 117-125.
Lasswell, Harold D. (1948). *Power and personality*. London:Chapman & Hall.
Lidén, Erik R. (2006) 'Stock Recommendations in Swedish Printed Media: Leading or Misleading?', *The European Journal of Finance*, Vol 12 (No.8), 731-748
Mandelbrot, Benoit. (1963). 'The variation of certain speculative prices'. *Journal of Business*. Vol. 36, pp 394-419.
Namenwirth, Zvi., and Lasswell, Harold D. (1970) *The changing language of American values: a computer study of selected party platforms*. Beverly Hills (Calif.): Sage Publications.
Stone, Philip J.,Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie et al (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Boston: The MIT Press. (For downloading the GI Lexicon, please go to http://www.wjh.harvard.edu/~inquirer/ ,Accessed 17th April 2008)
Taylor, Stephen J. (2005). *Asset Price Dynamics, Volatility, and Prediction.* Princeton and Oxford: Princeton University Press
Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, Vol. 62 (3), pp 1139-1168.

Annotating Opinion – Evaluation Of Blogs

Estelle Dubreil, Matthieu Vernier, Laura Monceaux, Béatrice Daille
LINA CNRS UMR 6241
2, rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, FRANCE
E-mail: Estelle.Dubreil, Matthieu.Vernier, Laura.Monceaux, Beatrice.Daille@univ-nantes.fr

Abstract

This paper deals with annotating opinions on a non-specific corpus of blogs. This work is motivated by a more general aim of building a generic method for detecting opinions. In accordance with this aim, we propose a linguistic model for the description of the opinion expression phenomenon.

1. Introduction

Up to now the previous approaches of sentiment analysis can be separated into two categories. The first one aimed to categorized texts according to semantic orientation (Turney 2002; Pang and al 2002; Gamon and al 2005; Torres-Moreno and al 2007) and most of the time we distinguish between the positive orientation or the negative one, even neutral.

The second approach concentrated his effort on the extraction of evaluations by taking a particular attention on the conveyed information. For example, to determine who think what about which product? By this way we find a richer and more precise information (Yi and al. 2003; Hu and al. 2004; Popescu and al. 2005; Kobayashi and al. 2007).

This paper present an annotating work based on a linguistic model to prepare such an extraction of evaluations approach. We focus on the different modality (*opinion*, *judgement*, *agreement* ...) of the evaluation expression on a corpus of weblog posts. Such corpus, also studied in (Chesley and al. 2006; Mishne and al 2006) is likely to be representative of the various language phenomenon linked to the expression of evaluation. By this study, we tend to develop a generical method in order to detect evaluation expressions on every kind of evaluated subject existing on blogs.

2. Blogosphere

2.1 Issues of an automatic detection of opinions

Blogs are everywhere nowadays: the media uses and refers to them, politicians resort to them and researchers use them for their work. With this popular fashion, the number of Blog platforms has increased in France since 2002, so multiplying tenfold the possible extent of information exchanges. On average 1100 Blogs are created every day on the platform 'Over-blog'[1], the French website form from which the corpus is extracted. Every visitor spends approximately 12 minutes of his or her day on its 3,5 million pages. They are read by a Blogger population that is representative of the global population, because Over-blog is not specifically marketed for teenagers, unlike the Skyblog and MSN Spaces for example. Blogs therefore constitute both a new method of information exchange and a new power of information, which can influence the opinion of readers. Therefore, Blogs represent an ideal object of study for the observation of different forms of expressions of opinion – evaluation, where evaluation includes all types of opinion. Representing a subset of the Web, Blogs are « mainly made up of published posts which are deposited and appear in a non chronological order (the most recent are found at the top of the page), and they often include external hypertexts or links » (Fievet & Turrettini 2004, p. 3).

2.2 A new media

This new form of media is specific because of the abolition of the impersonal character of communication. Every reader has the possibility of answering the post published by the author of a Blog by posting his or her own comment. In addition, as the publication of the information is becoming increasingly easy, this type of media has become very popular. The interactive specificity of Blogs is based on enunciation rules and the management of the relation with the public. This has resulted in the internet community rapidly adopting Blogs as a means of expressing favourable and unfavourable private states. If the expression of a kind of evaluation is the common denominator to all Blogs, the landscape of the Blogosphere is not homogeneous, but articulated around an axis which stretches from the personal diary to pure weblogs (link archives), by way of the thematic blog.

2.3 The personal thematic blogs

We worked on the personal thematic blogs written in French by adults, defined as " sites where cultural productions are critically evaluated […], because the enunciator declares him/herself initially as an individual capable of judgement and analysis, and who can evaluate and exchange on the subject of public objects " (Cardon et Delaunay-Teterel 2006, p.40).

Blogs represent a new kind of type of text which concentrates a large diversity of subjects, lexicon, morpho-syntactic patterns, which made that we are not limited to a specific domain. The subjects on which we are working on are as different as Harry Potter, Wii, Vladimir Poutine (Named entities), strike, ecology

1 Over-blog – http://www.over-blog.com/

(concept).
This work presents the annotation of a corpus of posts and their comments with the aim of developing a blog monitoring tool which is able to automatically detect the evaluation of bloggers with regard to a given entity. This manual labelling was made by tagging XML via the creation of a DTD which contains the notions of « subject » and « evaluation ». As detailed below.

3. Description of evaluation

3.1 Subject

The notion of "subject" contains two categories: the *concepts* and the *named entity*.
A concept is generally a noun or a nominal group, occurring in the text, and which is representative of the theme of the post or the comment which has been tagged. We defined three types of concepts: the *Concerned Concepts* - CC corresponding to the nature of the referent and answering the question " What is the post or the comment about ? ", the *Associated Concepts* - AC associated in field of the CC, and the *Non associated Concepts* - NC which contain an evaluation independent to the general theme of the post or the comment (Ex: post " Sin City ": CC: info, film, upload ; AC: date de sortie/release date, réalisateur/producer ; NC: ville/town, quartier/area, flics/police).
A named entity is generally a proper noun (Ex: Sarkozy, Weeds), occurring in the text, which inevitably contains an evaluation. We defined two types of named entity, the *associated named entity* – IA, associated with the semantic field of the post or the comment, and the *non associated named entity* – IN, which is not associated with the semantic field of the post or the comment.

3.2 Linguistic model

The notion of *evaluation* brings together linguistic data which can be observable through the phenomena of modalisation described by Charaudeau (Charaudeau, 1992), among which we extract the descriptions specified for the elocutive modality of enunciation (dedicated to the speaker), which he put to use in his study on the film criticism (Charaudeau, 1988). The choice of this linguistic model is made legitimate by the descriptions of the evaluation proposed, firstly because it enables researchers to find answers to traditional questions of this type (Schröder et al., 2006): "What can be qualified as evaluation? ", " How can we rank evaluations according to their positive or negative polarity ? ", " How does the context change the polarity and the strength of evaluation ? " Secondly, these descriptions are significant, because they come into being through a list of occurrences, and are important benchmarks, as was shown by the procedure of inter-annotations by Wiebe et al. (Wiebe et al., 2006).
Concretely, the elocutive modality of enunciation is divided into twelve modalities (Charaudeau, 1992). Five of these modalities (opinion, appreciation, acceptance-refusal, agreement-discord and judgment) refer to a type of evaluation: Each of these modalities contain many subcategories, also defined and clarified. For example, 'opinion' – OP is divided into five subcategories:

- *conviction* (Ex : je suis persuadé) ; I am absolutely sure
- *supposition certitude high* (Ex : je me doute) ; I am almost sure
- *supposition certitude medium* (Ex : je crois) ; I think
- *supposition certitude low* (Ex : je doute) ; I am not sure
- *supposition premonition* (Ex : je sens). I feel

In the same way, 'appreciation' is divided into six subcategories:

- *explicit appreciation favourable* – EAF (Ex : je suis satisfait) ; I am satisfied
- *explicit appreciation unfavourable* – EAU (Ex : je suis triste que) ; I am sad that…
- *explicit appreciation exclamative form favourable* – EAEF (Ex : Youpi!) ; I'm very happy !
- *explicit appreciation exclamative form unfavourable* – EAEU (Ex : Merde!) ; Dam !
- *implicit appreciation favourable* – IAF (Ex : c'est vraiment intéressant) ; This is very interesting
- *implicit appreciation unfavourable* – IAU (Ex : c'est vraiment mauvais). This is really bad

4. Data : Blogoscopie Corpus

The Blogoscopie Corpus was annotated according to the annotation scheme described above. At present it contains 200 posts and their comments, which represents 83500 words, extracted in June 2007 among 33 of 43 themes proposed on the website, and selected in accordance with personal thematic blogs (Ex: current events, blogzines, business, cinema, consumption/buying, beliefs, music, politics, etc.). With an aim to representing of the interests of bloggers, this extraction focused on the 10 most visited blogs according to each theme, and more particularly the first 10 posts published and their comments (maximum 10). The work of annotation is finished and the corpus will be made public at the end of 2008.

5. Annotating methodology

The annotation of 200 posts was carried out in four phases: the application of the linguistic model phase, the confrontation of the data phase, the consolidation of the annotation scheme phase, and the increase of the volume of data phase. The annotation instructions do not specify either the formal criteria for the identification of the various forms of evaluation, or the type of words to be annotated (Ex: verbs or adjectives, parts of speech, word classes). However the annotation does take into account the syntagmatic level.

5.1 Application of the linguistic model

The application phase for the linguistic model was simultaneously carried out by 4 annotators on 12 posts, chosen among very different themes for their linguistic difficulties. The aim was to evaluate the relevance of the annotation scheme so that common rules of annotation could emerge.

Regarding the entities, among the agreed inter-annotators principles, the most important entities are the annotations of all the AC even if they do not contain evaluation and, secondly, the annotations of the IA, only if they are the object of an evaluation. Next, if an evaluation has an impact on several entities, they have to all appear in the attribute "form"; the separator being the comma.

Regarding the evaluations, it seemed necessary to annotate the polarity according to the context of enunciation, to annotate the phrase logic forms (Ex: bailler aux corneilles/dead tired, regretter amèrement/bitterly regret), and to deconstruct the evaluations in case of a conjunction of modalities (Ex: cette *femme* [CC] **brillante** [*implicit appreciation favourable*] et **ravissante** [*implicit appreciation favourable*]/ this brilliante and charming woman).

5.2 Confrontation of the data

The confrontation phase of the data aimed to stabilise the annotation rules updated for the concepts. A team of 3 annotators were in charge of annotating concepts only on 64 posts. The annotation was carried out in three stages. Every post was annotated by two linguists, working individually. With a vantage to objectivity, a harmonisation stage was carried out by the third linguist, who had not participated in the annotation. This refined the annotation principles. From now on, when a text contains a spelling variant or a synonymic occurrence of a concept, it has to be annotated with the same identifier (Ex: série, séries, sitcom). When necessary, the AC are annotated as hyponymous of CC (Ex: *fruits secs/ dried fruit* [CC], *raisins/grapes* [AC]). Finally, the annotation of the concepts is made on the longest nominal group in the presence of a preposition (Ex: feuille de laurier/ laurel leaf, or in French ‘ leaf *of* the laurel). The first annotation showed great differences between the annotators. The first one considered in average 1.5 CC (up to 5) and 9 CA (up to 42) by post. The second one indicated 1.5 CC (up to 7) and 6.7 (up to 25) CA by post. The agreement between annotators was low: only 50% for the CC and 44% for the CA.

5.3 Consolidation of the annotation scheme and increase of the volume of data

The consolidation phase of the annotation scheme aimed to stabilize the annotation rules updated for the named entity and the evaluations. The annotation was made by a single annotator, but every question or problem was the object of a discussion with 4 other linguists and computer specialists. This work also allowed us to refine the annotation principles. For example, it seemed necessary to create an attribute "irony" (Ex: Ah bon ? parce que **ça marche moyen/ Really? Because it doesn’t really work** [*implicit appreciation unfavourable*] entre *Nico* [IA] et *Cécilia* [IA] ? Quelle blague!/ what a joke!), to integrate personal pronoun subjects into the prospect, so that they could be used later as markers of intensity (Ex: **Je suis persuadé que / I am sure that**[opinion conviction]), so as to annotate the exclamatory forms of the appreciation. Even when the exclamation mark is distant or even missing (Ex: **Contente/Happy** [*explicit appreciation exclamative form favourable*] de t'avoir *fait rire / to have made you laugh!* [NC]!!!). The new rules allow an improvement in these figures with an agreement of 63% for the CC and 52% for the CA.

The increase of the volume of data phase aimed to confirm the general annotation principle. The annotation was made by a single annotator on 124 blogs.

6. Statistics and observations

A part of our corpus is constitued by 100 blog pages talking about different kind of evaluated subject : a movie (“Le coeur des hommes 2”), a wine (“Beaujolais”), a person (“Raymond Domenech”), a social event (“The strike”), a book (“Harry Potter”), a polemical french law (“LRU”), an object (“Wii”) and two conceptual subject (“Sustainable development” and “Nuclear”). Tab bellow presents the repartition of the evaluation modality observed after the annotating process in these weblogs posts and in theirs comments.

	OP	*EAF*	*EAU*	*IAF*	*IAU*
Eval. markers in posts	68 5.84%	43 3,7%	19 1,6%	434 37,25%	472 40,52%
Eval. markers in com-ments	22 6.5%	31 9,1%	8 2,4%	103 30,3%	131 38,5%
Movie	14	15	4	51	32
Wine	4	4	2	98	31
Person	12	9	6	80	115
Social event	8	4	1	32	57
Book	7	25	5	72	33
Law	24	8	5	84	155
Object	5	6	2	32	31
Sus-tainable develop-ment	5	3	2	57	49
Nuclear	11	0	0	31	100

Table 1: Repartition of the evaluation modality.

Through these different kinds of subjects, we observe regularities used to formulate an evaluation. Proportion of modalities seems to be separate from the analysed subject. In this way, we notice that “Implicit appreciation” modality is the most frequently used by bloggers. Next, we find more explicit modality of evaluation by the using of phrase indicating a “judgement” or an “opinion”. On an other hand, according to analysed subject, we can already note which subjects have a favourable evaluation and those who don’t.

All regularities in the observed modalities and the fact

that these regularities are separated from the kind of subject treated will be used to automatically detect evaluations in new weblogs and subjects of any kinds.
At the same time, first all kind of the annotated evaluations were got back to establish lexical resources, which specified the positive or negative value of the data. Secondly, we study the relation between an evaluation and the evaluated subject.
For example:
Vladimir Poutine, qui jouit d'une forte popularité en Russie/Vladimir Poutin, who enjoys a strong popularity in Russia.
<IA cc="Président, Russie">**Vladimir Poutine**</IA>, **qui** <Appreciation type="IAF" subject="Poutine">**jouit d'une forte popularité**</Appreciation> en Russie

Harry Potter 3 est une réussite totale/ Harry Potter 3 is a total success.
<IA cc="film, saga">**Harry Potter 3**</IA> <Appreciation type="IAF" forme="Harry Potter 3">**est une réussite totale**</Appreciation>

We also establish automatically a list of regular morpho-syntactic patterns as:

- [Subject] + qui + [evaluation] (cf. beside) ;
- [Subject] + être + [evaluation]

All these patterns confine on the intra-phrastic level. Moreover by using intra-phrastic patterns, (Kobayashi and al., 2007) showed that it was possible to find a relation subject-evaluation in new posts with a precision of 0,56 and a recall of 0,53. These first observations tend towards a sketch of a grammar of the expression of the evaluation in blogs, just like the specifications obtained by Legallois and Ferrari (Legallois & Ferrari, 2006) on of cultural objects evaluations. These regularities will be automatically found in new posts by a bootstrapping method (Riloff & Wiebe, 2003).

7. Conclusion

Starting from a linguistic model for the description of evaluation, we create an annotation system enable to list the specific linguistic marks resulting from the expression of evaluation in blogs. This annotation scheme distinguishes between the subject and the evaluations concerning these subjects. By this way, we applied this scheme to posts and comments. Also, we have henceforth a corpus under format xml constituted by 200 annotated blogs and of their associated comments, what represents 83500 words. Besides a stabilized annotation scheme, we also have rules of annotation the relevance of which was validated, by an inter-annotators agreement. The first observations led on the corpus show that the proportion of modalities seems to be separate from the analysed subject, because we notice the most important frequency of the implicit appreciation compared to the other expression of evaluation. At the moment, we build first a lexical resource, which specified the positive or negative value of the data, and a list of regular morpho-syntactic patterns, for tending towards a sketch of a grammar of the expression of the evaluation, which will be automatically found in new posts by a bootstrapping method.

8. References

Cardon, D., Delaunay-Téterel, H. (2006). *La production de soi comme technique relationnelle : un essai de typologie des blogs par leurs publics*. Les blogs, Réseaux Vol. 24 N° 138/2006, pp. 15-71.

Charaudeau, P. (1998). *La critique cinématographique : faire voir faire parler*. Dans La Presse, Produit, Production, Reception, Paris, Didier Erudition, pp. 47-70.

Charaudeau, P. (1992). *Grammaire du sens et de l'expression*. Paris : Hachette éducation, 925 p.

Chesley, P., Bruce, V., Li, X., Rohini, S. (2006). *Using Verbs and Adjectives to Automatically Classify Blog Sentiment*. To appear in AAAI Spring Symposium Technical Report SS-06-03.

Fievet, C., Turrettini, E. (2004). *Blog Story*, Eds. Eyrolles, 305 p.

Gamon, M., Aue, A. (2005). *Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms*. In Proceedings of the ACL-05 Workshop on Feature Engineering.

Hu, M., Liu, B. (2004). *Mining and summarizing customer reviews*. Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining (KDD).

Kobayashi, N., Kentaro, I., Matsumoto, Y. (2007). *Extractingaspect-evaluation and aspect-of relations in opinion mining*. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Legallois, D., Ferrari, S. (2006). *Vers une grammaire de l'évaluation des objets culturels*. Schedae, Prépublication n° 8 (fascicule n°1, pp. 57-68).

Mishne, G., Glance, N. (2006). *Predicting movie sales from blogger sentiment*. AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006).

Pang, B. Lee, L., Vaithyanathan, S. (2002). *Thumbs up?* Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 79–86.

Popescu, A. Etzioni, o. (2005). *Extracting product features and opinions from reviews*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Schröder, M. Pirker, H., Lamolle, M. (2006). *First suggestions for an emotion annotation and representation language*. In L. Deviller et al. (Ed.), Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect (pp. 88-92). Genoa, Italy.

Torres-Moreno, J.-M., El-Bèze, M., Béchet, F., Camelin, N. (2007). *Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ?* Application au défi fouille de textes 2007. DEFT07, pp119-133, AFIA 2007, Grenoble, France.

Turney, P.D. (2002). *Thumbs up or thumbs down ? semantic orientation applied to supervised*

classification of reviews. In 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia.

Wiebe, J., Wilson, T., Cardie, C. (2005). *Annotating Expressions of opinions and Emotions in Language*. Language Ressources and Evaluation, Vol. 39, issue 2-3, pp. 165-210.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. (2003). *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques*. Proceedings of the third IEEE International Conference on Data Mining (ICDM).