Preface

Although it is generally assumed that improvements in language processing will be made through the integration of linguistic information and statistical techniques, the reality is that language is very diverse and looking for specific patterns of words that repeat enough to be statistically significant tends not to be a very fruitful task: sequences longer than three words are not generally repeated often enough to be statistically significant. At the same time, the identification of named entities: names, dates, places, organizations etc., has proved to be a very useful preliminary task in many natural language processing systems. This workshop is dedicated to the discussion of approaches which extend this notion by identifying and labeling other semantic information in a text, in such a way as to allow repeatable semantic patterns to emerge. The papers selected focus on ways to attack the data sparseness problem by collapsing (semantically) related phrases which are expressed by different word sequences.

As this seems closely related to previously proposed class-based language models (see for example Brown et al. 90 in Computational Linguistics), it is different in that the empirical notion of classes used in the previous work (e.g. classes made up of collocationally similar words) are replaced by semantically justified sets.

Notice how Name Entity (NE) tagging and Word Sense Disambiguation (WSD) represent, in terms of granularity and representational complexity, two extremes of a single general problem: semantic disambiguation. Semantic disambiguation serves thus the purpose of improving the generalization power of statistical models. One of the questions here is how to determine a suitable level of clustering (for NE identification and for WSD) that would lead to high accuracy and to performance improvement by obtained statistical models.

It is to be noticed that several independent research efforts that focused recently on the statistical treatment of semantic phenomena (e.g. WordNet navigation as a stochastic process, as studied in Light and Abney or in Ciaramita & Johnson, 2003) correlated highly with the research program proposed above.

The workshop will offer a forum where experience from lexical semantics and statistical learning will be presented and fruitful discussion among researchers in both fields will be promoted. The workshop is expected to attract researchers and practitioners from a range of areas as well as developers of large scale semantic resources who are interested in effective methods of semantic labeling.

The main topics of the workshop can be not exhaustively listed as follows:

- Methods for lexical - semantic annotation of corpora
- Methods and standards for lexical semantic representation of dictionary information
- Lexico-semantic taxonomies
- Existing sources of classification: dictionaries, thesauri and computerized ontologies
- Corpus-driven methods for semantic disambiguation
- Feature selection for semantic disambiguation
- Lexico-semantic tagging of very large corpora
- Algorithms and methods for disambiguation of semantic phenomena
- Statistical learning models and their applications to semantic labeling
- Computational learning frameworks for Natural Language Learning
- Semi-supervised and unsupervised statistical semantic disambiguation
- Evaluation of semantic disambiguation

Technical papers gathered in the Proceedings represent a specific contribution to the above complex issues.

April, 20st 2004
Louise Guthrie, Roberto Basili , Eva Hajicova, Frederick Jelinek

Program Chairs
LREC04 Workshop on
”*Beyond Named Entity Recognition: Semantic labelling for NLP tasks*”

Workshop Programme

8:45-9:00 *Welcome*

9:00-9:20 *Reducing the effect of name explosion*
Dimitrios Kokkinakis, Department of Swedish Language, Sprkdata, University of Gothenburg

9:20-9:40 *The UCREL Semantic Analysis System*
Paul Rayson, Dawn Archer, Scott Piao and Tony McEnery, UCREL, Lancaster University

9:40-10:00 *Towards filling the gap between lexicon and corpus*
Iulia Nica, M. Antonia Marti', (CliC, University of Barcelona, Spain)
Andrez Montoyo, Sonia Vazquez, (Departmento di Linguistica Generale, University of Alicante, Spain)

10:00-10:20 *Semantic Annotation of Multilingual Text Corpora*
Teruko Mitamura (Carnegie Mellon University), Keith Miller (MITRE Corporation),
Bonnie Dorr (University of Maryland), David Farwell (New Mexico State University),
Nizar Habash (University of Maryland), Stephen Helmreich (New Mexico State University),
Eduard Hovy (University of Southern California), Lori Levin (Carnegie Mellon University),
Owen Rambow (Columbia University), Florence Reeder (MITRE Corporation),
Advaith Siddharthan (Columbia University)

10:20-10:40 *Coffee Break*

10:40-11:00 *Verb Classification Machine Learning Experiments in Classifying Verbs into Semantic Classes*
Bart Decadt and Walter Daelemans
Center for Dutch Language and Speech (CNTS), University of Antwerp, Belgium

11:00-11:20 *Unsupervised Semantic Tagging*
Roberto Basili, Marco Cammisa, Department of Computer Science, University of Roma, Tor Vergata, Italy

11:20-11:40 *A WordNet-based Algorithm for Unsupervised Relation Extraction*
Mark Stevenson, Department of Computer Science, University of Sheffield

11:40-12:00 *Exploiting the Semantic Fingerprint for Tagging Unseen Words*
Fabio Massimo Zanzotto, Armando Stellato
Department of Computer Science, University of Roma, Tor Vergata, Roma, Italy

12:00-12:20 *Extended Semantic Tagging for Entity Extraction*
Narjes Boufaden, Guy Lapalme, Yoshua Bengio, Universite' de Montreal, Quebec, Canada

12:20-13:00 *Round Table*
chairs:
Eva Hajicova (Charles University, Czech Republic),
Frederick Jelinek (Johns Hopkins University, Maryland, USA)

Organization

The Workshop is organized as a joint cooperation between the University of Sheffield (UK), the University of Roma, Tor Vergata (Italy), the Charles University of Prague (CZ) and the Center for Language Sand Speech Processing of the John Hopkins University (MD,USA).

Program Commitee

Louise Guthrie (University of Sheffield, UK)
Roberto Basili (University of Roma, Tor Vergata, IT)
Eva Hajicova (Charles University, Czech Republic)
Frederick Jelinek (Johns Hopkins University, Maryland, USA)

Contents

Author Index

Reducing the effect of name explosion

Dimitrios Kokkinakis

Språkdata, Department of Swedish Language
University of Gothenburg, Box 200
SE-405 30, Sweden
dimitrios.kokkinakis@svenska.gu.se

Abstract

The problem of new vocabulary is particularly frustrating once one begins to work with large corpora of real texts. The identification of unknown proper nouns, chains of non-proper nouns and even common words that function as names (i.e. named entities) in unrestricted text, and their subsequent classification into some sort of semantic type is a challenging and difficult problem in Natural Language Processing (NLP). Systems that perform Information Extraction, Information Retrieval, Question-Answering, Topic Detection, Text Mining, Machine Translation and annotation for the Semantic Web have highlighted the need for the automatic recognition of such entities, since their constant introduction in any domain, however narrow, is very common and needs special attention. Proper names are usually not listed in defining or other common types of dictionaries, they may appear in many alias forms and abbreviated variations, which makes their listing infeasible. This paper deals with some extensions to the "traditional" named entity recognition approaches. It puts emphasis on more name classes and their further subclassification into finer sets. An operative system that can be tested and evaluated on-line implements the ideas described in this paper.

1. Introduction

There is a prevailing consensus between Natural Language Processing (NLP) practicioners that the possibility of achieving significantly better performance on natural language tasks requires knowledge-based processing, which can only be achieved by enhancing and using knowledge bases, ontologies, semantic lexicons and thesauri. The idea behind semantically annotating, e.g. words, with such resources is that some automatic process may use the markings added, in order to choose the proper concept underlying the words in a given context and thus get closer to a deeper, semantic disambiguation and understanding of the discourse in question. A syntactic parser, for instance, without accessing semantic information of this kind, will take us only part of the way, while a combination of syntax-semantics has better prospects. Better parsing results can be achieved if the semantics of each or at least some of the lexical items (e.g. heads of NPs) could be pre-determined, and thus aid a parser in constructing a more semantically-oriented phrase structure for a given sentence. This paper deals with an approach towards this goal. It describes the creation of an elaborated named entity hierarchy and its implementation into a named entity recognition system. The motivation behind this work has been that (newspaper) texts contain a plethora of names (often in long chains) that are not covered by existing name schemes, and consequently cannot be resolved or characterized by current word sense disambiguation (the ultimate goal) or named entity recognition systems. Thus, some way of proceeding towards the direction of defining, implementing and actually using larger and finer-grained classification schemes for names should be a step closer to WSD and thus natural language understanding. In this paper we will focus on the development, implementation and use of an enhanced version of such a NER system and its application on general Swedish corpora.

2. Background

Named entity recognition (NER), semantic tagging (ST) and word sense disambiguation (WSD) are related techologies that aim at the resolution of lexical ambiguity, either on a smaller scale (named entities), or on a larger scale (all the content words in a text). These technologies occupy a continuoum in terms of granularity of the semantic disambiguation problem, the first and third being the two extremes of it. For WSD for instance, the initial, and coarsest level of disambiguation would be just performing homograph distinction of typical verb-noun ambiguities, such as between 'play' as noun and verb. Semantic tagging, on the other hand, is defined as the more general instance of the lexical ambiguity problem, in which the labels assigned to the words in a text are broad semantic categories, or clusters of semantically related concepts.

Previous approaches in the field of NER include the well known MUC exercises in limited, and well defined domains (Grishman and Sundheim, 1996), the IREX initiative (Sekine and Isahara, 2000), and the most current ACE effort (EDT, 2000). All approaches have used limited named entity sets. MUC recognized not more than seven types of named entities, 'organization, location, person, date, time, money and percent expressions'. In IREX and in the Concerto project (Black et al., 2000), another kind of named entity was added to the MUC set, namely 'artifact'. While in the ACE, two new entities, 'geo-political entity' and 'facility', were added to pursue the generalization of the technology – two entities that were subsumed by 'location' and 'organization' in MUC.

However, in general language, as found in newswire and newspaper texts, many more types of "names" or "named entities" are likely to be encountered, and thus finer distinctions and more detailed level of analysis are required in establishing a more stable, robust and elaborate hierarchy for the subsequent recognition and annotation tasks, a crucial step for a number of NLP technologies (e.g. IE, IR, Q&A, TM, MT, annotation for the Semantic Web). One of the most ambitious projects w.r.t. a NE hierarchy was DR-LINK (Paik et al., 1996). They considered nine branching, 'geographic entity, affiliation, organization, human, document, equipment, scientific, temporal and misc.', and 30 terminal nodes.

However, labels such as ‘document’ seem too restricted and isolated, compared to the more general ‘human’ and ‘geographic entity’, while ‘affiliation’ and ‘organization’ seem to be in conflict with each other. In a similar fashion, Sheremetyeva et al. (1998) present an attempt to create a multilingual onomasticon with five top-level categories ‘occasion, animate, artifact, place and organization’ and 45 semantic categories in total. Finally, Sekine et al. (2002) reported the design and the development of a NE hierarchy with 150 types, organized in a tree structure. In the same paper, Sekine et al. argue that the definition of what is a NE is ambiguous and once we include artifact names (names of classes and not specific individuals) we might have a problem. However, artifact names and other proper name classes, are certainly a benefit for NLP applications, if these are identified as named entities. Thus, by accepting this argument the next step is to decide how far we want to go into the class to be named entity. The border between proper names and named entities is unavoidably ambiguous, and some arbitrary decision is necessary, even in our case.

2.1 So, what is meant by named entity?

Before we go into details on our system and name hierarchy, a couple of words on the key-term itself should be in place. Defining what a Named Entity (NE) is, is not a trivial enterprise. Simplistically, NEs are oftenly considered to be proper names occurring in texts. However, NEs go beyond what traditional grammar calls proper names or proper nouns (names of unique individuals or group of individuals). Even the difference between the last two is not quite clear, but since the former term (proper names) avoids specifying the part of speech of the linguistic unit of interest and since these entities are usually but not exclusively nouns, it is commonly used to refer to all single and multiword NEs.

NEs can have substantial internal structure; for instance, common nouns may form part of the proper name, “New York City”; a proper name may consist entirely of common nouns, “The Institute of Arts”; prepositions, articles and conjunctions may form part of the name, “University of California”; names may contain their description “the World Intellectual Property Organisation” etc. Furthermore, a single proper name may consists of all capital letters “EMU”, a mixture of capital and lower case letters “GlaxoSmithKline”, letters and digits, “JAS-39”, only digits “3” etc. Apart from syntactic characteristics, the meaning associated to a proper name can depend on personal experience of individual speakers and might not be constant across the majority of languages users in the way that lexical meaning can be expected to be, which complicates the issue of what can or cannot be annotated as NE. In some cases, meaning appears to be an optional element of proper name content which doesn’t seem to be the case with other vocabulary. Proper names may mean but do not have to while it is a neccessary property for other lexical items, which always have to carry some meaning; examples of such proper names are the case of cryptic designations such as “[supernova] SN 1987 A” or “[aircraft] F117A”.

Note, that in Information Extraction, and in the Named Entity Recognition task in particular, typical proper names and the expressions they form (e.g. organisations) are not the only element of interest. Numerical expressions (e.g. percentages) as well as time and date expressions share almost equally status and also treated as NEs. Obviously, such categories cannot be characterised as proper names; however, they share many of the characteristics that are outlined previously.

3. The Swedish NER-system

The Swedish system has been developed within the Nomen Nescio (NN) project. The NN-project is a Nordic Research Council (NORFA) financed network, within the language technology field, that deals with the recognition, classification and annotation of names in running text for three nordic languages (Swedish, Norwegian –bokmål– and Danish). The NN project presupposes no particular applications and therefore considers the task of name recognition from the point of view of general texts, particularly newspaper and scientific articles, novels etc. The Swedish system handles 8 main types, and 47 subtypes of NEs, and it does so in a modular and scalable manner[1]. The system consists of five major components, making a clear separation between lexical and grammatical resources, e.g. lists of multiword names, single names, grammars and algorithms. From own experience, we are convinced that no individual criterion (e.g. lists of names, grammars) can achieve both high precision and high recall. Therefore, we have defined and used a combination of criteria (external and internal evidence[2]; *cf.* McDonald, 1996) that together support effective identification of the target names. The five components are:

- lists of multiword names (approx. 3 000) particularly for locations and organizations, (MWE); e.g. “Amerikanska Jungfruöarna” “New York Rangers”, “ITTF Pro Tour”, taken from various Internet sites

[1] At the initial stages of the system development there was a thought of using as starting point part-of-speech annotated material. This would have certain advantages, such as using a simple label for a shorthand for all possible prepositions, but it would also have a number of disadvantages as well. For instance, that it would always require a part-of-speech tagger or part-of-speech tagged material in order for the user to test the functionality of the system. Thus, the portability of the recognizer would decrease dramatically since each new user would not only be dependent of a part-of-speech tagger, but also of a tagger that can produce a particular type of pos-annotation. Therefore, we abandoned the idea of implementing a system dependent on grammatical tagging and we chose to tailor the system in a way that the only requirement would be to have tokenised input.

[2] According to McDonald (1996) names, in spite of their diversity, have a systematic and compositional structure that can be captured using context-sensitive grammars. The need for such grammars is due to the fact that the classification of names involves two complementary kinds of evidence, *internal* and *external*. Internal evidence is taken from within the sequence of words that comprise the name, such as the content of lists of proper names (gazetteers), abbreviations and acronyms (Ltd, Inc., Gmbh). External evidence is provided by the context in which a name appears – the characteristic properties or events in a syntactic relation (verbs, adjectives) with a proper noun can be used to provide confirming or criterial evidence for a name’s category – a very important type of complementary information since internal evidence can never be complete.

- a shallow parsing component that uses context sensitive finite-state grammars, one grammar for each type of entity recognized (FSG); e.g. taken from the measure grammar "([0-9]+|[0-9]+/[0-9]+|[0-9]+,[0-9]+|[0-9]+-[0-9]+|...)" "(miljoner" "|...)?(g|kilo|kilogram|mikrog|kg|ton|TWh|...)" which matches strings such as: "100g" or "10,1 gram". Due to the spraseness of relevant examples for some of the groups, some of the rules have been manually written, others have been written by examining morphosyntactic annotated corpora, some have been supported by lexical information from defining dictionaries (e.g. typical predicates for various groups. "säga [say]" is a typical verb for humans, "jama [miaow]" is for animals and "storma [escalade]" for functional locations) while some have been semi-automatically generated and completed with more data while testing on large corpora
- a module that uses the annotations produced by the previous two components (which have a high rate in precision) in order to make decisions regarding entities not covered by FSG or MWE. This module is inspired by the Document Centred Approach by Mikheev *et al.* (1999) and Mikheev (2000). This is a form of on-line learning from documents under processing which looks at unambiguous usages for assigning annotations in ambiguous words (DCA); e.g. "<ENAMEX TYPE="ORG" SBT="CRP"><METHOD MTH="*MWA*/>LGP Telecom</ENAMEX> redovisar en vinst […] <ENAMEX TYPE="ORG" SBT="CRP"><METHOD MTH="*DCA*"/>LGP</ENAMEX>bedömer att [...]"; which annotates the second occurrence of "LGP" as "ORG/CRP" considering that "LGP Telecom" is unambiguously annotated as such. This module is applied twice, once after the grammars have been applied and once at the end of the single names look-up
- lists of single names (approx. 95 000), i.e. gazetteers (GAZ)
- a theory revision and refinement module makes a final control on an annotated document with named-entities in order to detect and resolve possible errors and complete with missing information based on existing annotations. Furthermore, if one of the previous modules fails to generate an annotation for some examplar, this module can guess at a refinement or correction, allowing valid annotations, this time based on an extended, mixed context of previous annotations, a few trigger words and orthography. (E.g. given an annotation "<ENAMEX TYPE="ORG" SBT="CRP"> METHOD MTH="*GAZ*"/>Vodafone </ENAMEX>, Orange, <ENAMEX TYPE="ORG" SBT="CRP"> METHOD MTH="*GAZ*"/>T-Mobile </ENAMEX>" this module will guess that "Orange" is also an organization).

4. The NER Taxonomy

As can be seen from the way "names" are defined and used in MUC, the NE categories go beyond what traditional grammar define as proper names or proper nouns. The definition glides towards what seems to be a semantic categorisation in general terms. This is inline with the fact that semantic annotation is not as well understood as grammatical annotation, since there is no consensus on the tagset and its content to be used. In practice, semantic annotation is a compromise between attempts to mirror how words are related in the human mind, and the need for usable annotated corpora. Thus, using semantic classes for names we can capture high-level abstractions for surface words, and hence shift from looking at words to looking at groups or classes of words. The eight core categories distinguished in the Swedish system, with their names and associated tags are: location "LOC", person "PRS", organisation "ORG", event "EVN", object "OBJ", work & art "WRK" time "TME" and measure "MSR". Each core category, "TYPE", is further organized in a network of totally 47 subtypes, "SBT", details for all subcategories are given in the Appendix.

4.1 Annotation

The NER effort consists of a number of subtasks that correspond to a number of XML tag elements pointing to the beginning and end of the entity. We chose as the name of the core element the widely used in MUC and ACE "ENAMEX" tag. Moreover, we use two attribute names for each entity, namely its major type "TYPE" and its sub-classification, subtype, "SBT". Finally we use an extra empty tag "METHOD" with a single attribute "MTH" that denotes which technique has been used for the identification of the entity, that is FSG, GAZ, MWE or DCA (see previous discussion). Thus, an annotated entity may look like the following:

<ENAMEX TYPE="PRS" SBT="HUM"><METHOD MTH="GAZ"/>Kalle Svensson</ENAMEX>

The above can be interpreted as: "Kalle Svensson" is an entity of type person, subtype human, identified using the gazetteers method. For validation purposes an XML schema has been implemented covering the NE hierarchy.

5. Evaluation of the NER system

The Swedish NER system was tested on *newspaper texts* (articles from 7 daily newspapers), *scientific texts*, articles from four *women's magazines*, and two randomly selected *literary passages* from A. Strindberg's (chapter 1 of "Götiska Rummen" and chapter 3 of "Svenska Folket 2"). The texts were divided into eight content-related groups: *newspaper foreign news* (FRGN), *newspaper sport news* (SPRT), *newspaper cultural news* (CLTR), *newspaper economy news* (ECNM), *newspaper domestic news* (DMST), *written scientific texts* (SCIE), *women's magazines* (WOMN) and *literary* (NOVL). Each group (except the literary texts) consisted of 10 randomly chosen texts downloaded from Swedish Internet sites the 31st of Oct. 2003. The total number of tokens in the material was 45 962, while the manually annotated names were 2 147[3].

5.1 Evaluation Measures

The sample texts were evaluated according to *precision*, *recall* and *f-measure* according to the formulae given below. The calculation considers two parameters, the attributes TYPE and SBT (SUBTYPE) of the ENAMEX

[3] Person names, such as 'first-name last-name', or 'first-name middle-name last-name' are considered as *one* name.

annotation and the portion of the name-segment that is matched by the system. The metrics calculated solely on TYPE do not consider the SBT attribute. Partially correct answers are measured according to the following MUC-inspired penalty scheme. Penalty is a figure used when there are name segments not marked by the automatic process, given a manually recognized name segment; where the entity type and/or subtype is/are correct but the span is incorrect or when there is case of an ambiguous annotation[4]. The penalty is defined according to the portion of a segment matched or not matched. For instance, if less than 40% of name is correctly matched by the system the penalty is 0,25, if a 40-60% of a segment is correctly matched the penalty is 0,5 and if it is 60%-99% of a segment is matched the penalty is defined to be 0,75. The penalty figures are further divided by 2, in the case of a partial matched name in conjunction with an erroneous subtype annotation. In a case of an ambiguous tag with correct span the penalty is 0,25.

Precision = (Total Correct + (Penalty*Partially Correct)) / Total Returned
Recall = (Total Correct + (Penalty*Partially Correct)) / All Possible

F-value is defined as the ratio:
F-value = (2 * Precision * Recall)/(Precision + Recall)
Thus, the evaluation was perfomed using all of the system's resources on:

- the main type TYPE of the annotations (e.g. <ENAMEX TYPE="XXX">*Name*</ENAMEX>)
- both the main TYPE and subtype SBT of the annotations (e.g. <ENAMEX TYPE="XXX" SBT= "ZZZ">*Name*</ENAMEX>)
- the amount of the name matched (e.g. <ENAMEX TYPE="XXX" SBT="ZZZ">*Name1 Name2*</ENAMEX> vs. <ENAMEX TYPE="XXX" SBT="ZZZ">*Name1* </ENAMEX> *Name2*)

5.2 Average Scores

The results, separately for TYPE and TYPE&SUBTYPE, shown below are for all the name groups except "MEASURE" and "TIME", which are unproblematic and score on near human perfection.

ALL GROUPS			
P	R	f-score	based on
0,942	0,871	(2*0,942*0,871)/ (0,942+0,871) = 0,905	TYPE
0,935	0,865	(2*0,935*0,865)/ (0,935+0,865) = 0,898	TYPE+SBT

Figure 1: Average P&R&F-score

Figure 1 shows the precision, recall and f-score for all groups based both on the TYPE and the TYPE & SUBTYPE.

5.3 Discussion on the Evaluation

The results are consistently high, with percentages in the mid to high 90s, particularly for the six groups taken from newspaper sites. This is not surprising given that more time has been spent on testing the system to newspaper material, which has been the system's major source of knowledge. Somewhat surprising, were the "SCIE", popular science texts, which scored high, considering both the TYPE and the SUBTYPE attributes (P=98,7%, R=91,6%). The texts taken from women's magazine achieved the lower scores (P=90,4%, R=82,5%). This can be explained by examining the errors produced and the names missed by the system on this category. This was largely because of the fact that names were used without any 'typical' contextual clues. 17 of the errors have to do with names of contemporary films, e.g. five occurrences of "Down with love" and three of "Smala Susie".

How are the results measure[5] with NER systems in the international arena? In order to answer this question we have to consider that NER systems usually deal with names of persons, locations, organizations, time and numerical expressions and on rather narrow domains. Thus, making such a direct comparison will be unfair to the Swedish system. The closer we can get to such comparison is to consider the annotations of person, location and organization names of our evaluated material and attempt a rather coarse comparison with other systems. If we just look at the errors produced and the missed names for these three "generic" categories, we get the following results:

ERROR ANALYSIS (PERSON/LOCATION/ORGANIZATION)						
	prs-wrong	loc-wrong	org-wrong	prs-miss.	loc-miss.	org-miss.
ECNM	0	1	7	0	0	9
FRGN	0	2	0	3	1	2
SPRT	1	5	10	6	1	2
DMST	0	2	2	0	4	2
CLTR	2	4	2	2	5	6
SCIE	1	1	1	2	2	3
WOMN	25	6	0	4	12	7
NOVL	9	0	3	16	12	5
TOTAL	38	21	25	33	37	36

Figure 2

Considering that in the evaluation material there were 1 026 annotations for person, 597 for location and 305 for organisation, we can now get a rough estimate of the precision and recall for the three groups (see below). A comparison can be then made with the MUC6&7 results in which the f-score for MUC6 in the NER task was <97% and the f-score for MUC7 was <94%[6]. The average f-

[4] A few ambiguous cases are still allowed by the current version of the system – this will be eliminated in the next version.

[5] Several successful systems for large-scale NER have been constructed during recent years, ranging from manually created rule-based systems to fully automatic learning-based systems. The borders between the different approaches are rather fuzzy and comparison between the various methods should be made with caution, since it is difficult to decide whether the results reported are truly comparable with each other. Different researchers use different quantities of data for (sometimes) training, and evaluation, and they also often define the metrics of precision and recall in a slightly different manner.

[6] http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html. The results include TIMEX and NUMEX.

score for the three groups above results to 93,8%, which is quite in accordance with the last MUC results.

	P	R	F-score
PRS	1026-38/1026≈96,2%	(1026-38)/33+(1026-38)≈96,7%	2*P*R/P+R ≈96,4%
LOC.	597-21/597≈96,4%	(597-21)/37+(597-21)≈93,9%	2*P*R/P+R ≈95,1%
ORG	305-25/305≈91,8%	(305-25)/36+(305-25)≈88,6	2*P*R/P+R ≈90,1%

Figure 3

6. A Note on Metonymy

Metonymy (or regular polysemy, even called semantic ambiguity), the phenomenon that when a speaker uses a reference to one entity to refer to another entity – or entities – related to it; is *the* major headache for in NER. According to Lakoff & Johnson (1980) metonymy is a form of figurative in which one expression is used to refer to the standard referent of a related one. The reference to one entity is usually done explicit while the other indirect. A typical, frequent example in the newspaper material is the case of capital city names standing in for national governments. In some sense, *all* words are metonyms, and this is the case we have experienced in our work. At this stage, we haven't make any calculations regarding the percent of metonymy in the texts either per type of NE or totally, this is an issue we will investigate in the future. Most of the metonymies can be resolved using near context. For instance three of the "Volvo's" senses (the organisation, the vehicle and the share) can be captured by typical context words :

- "{koncernen, dominerar, tycker…} Volvo" denotes "<ENAMEX TYPE="*ORG*" SBT="*CRP*">" while
- "{röd, grön, blå, gul…} Volvo" denotes "<ENAMEX TYPE="*OBJ*" SBT="*VHG*">" and
- "{steg, backade…} Volvo" denotes "<ENAMEX TYPE="*OBJ*" SBT="*PRD*">".

7. Conclusions and Further Work

Named Entities occupy a considerable proportion in natural language and have remained an important, partly unexplored area in NLP. In this paper, we have presented an implemented and evaluated Swedish NER system using a rich name hierarchy, developed within the Nomen Nescio project. The evaluation texts chosen belong to widely different genres. The results are very promising, in that we are able to achieve high recall and precision scores, comparable to MUC7 results, despite the fact that we use more general corpora and larger set of named entities. To get a better understanding of the system's capabilities we need to continue development and testing, using in particular texts of different style and structure, e.g. emails, spoken/transcribed input etc. and make a study on the adaptations and modifications that are required for such input. Moreover, it is desirable to put the system in a larger perspective, probably use it in conjunction with a concrete application. A regression test is planned during 2004 in order to track the system's preformance over time, since changes made to the system might have implications over the evaluation corpus. It is not uncommon that improvements in one case can lead to problems in others. A natural extension to the NER system is to couple it with a co-reference resolution module. Furthermore, since the implemented system is highly modularized, the taxonomy of types can be restricted for specific domains. In this direction there is some planned work for the anonymisation of electronic patient/health records within the EU-funded SemanticMining project (6th framework). The system can be tested and evaluated on line at: http://g3.spraakdata.gu.se/nn/.

Acknowledgments

Part of this research is supported by the NORFA (www.norfa.no) and particularly the Nordic research programme for Language Technology (2000-2004).

References

ACE [EDT] (2000). *Entity Detection and Tracking – Phase 1*. Doc: edt_phase1-v2.2. ACE homepages: http://www.ldc.upenn.edu/Projects/ACE2/ and http://www.itl.nist.gov/iaui/894.01/tests/ace/phase2/index.htm (visited Oct. 2003).

Black W.J., McNaught J., Zarri G.P., Persidis A., Brasher A., Gilardoni L., Bertino E., Semeraro G. and Leo P. (2000). A Semi-automatic System for Conceptual Annotation, its Application to Resource Construction and Evaluation. Proceedings of the *Second Language Resources and Evaluation Conference (LREC)*. Athens, Greece.

Grishman R. and Sundheim B. (1996) Message Understanding Conference - 6: A Brief History. Proceedings of the *16th International Conference on Computational Linguistics*. Copenhagen, Denmark.

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

McDonald D. (1996). Internal and External Evidence in the Identification and Semantic Categorisation of Proper Nouns. *Corpus-Processing for Lexical Acquisition*, 21-39. Pustejovsky J. and Boguraev B. (eds). MIT Press.

Mikheev A., Moens M. and Grover C. (1999). Named Entity recognition without gazetteers. Proceedings of the *EACL'99*. pp. 1-8. Bergen, Norway.

Mikheev A. (2000). Document Centered Approach to Text Normalization. Proceedings of the *SIGIR '2000*. pp. 136-143. Athens, Greece.

Paik W., Liddy E.D., Yu E. and McKenna M. (1996). Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval. *Corpus Processing for Lexical Acquisition*, Boguarev B. and Pustejovsky J. (eds), pp. 61-73, Bradford.

Sekine S., Sudo K. and Nobata C. (2002). Extended Named Entity Hierarchy. Proceedings of the *Third Language Resources and Evaluation Conference (LREC)*. Las Palmas, Spain.

Sekine S. and Isahara H. (2000). IREX: IR and IE Evaluation project in Japanese. *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*. Athens, Greece.

Sheremetyeva S., Cowie J., Nirenburg S. and Zajac R. (1998). A Multilingual Onomasticon as a Multipurpose NLP Resource. Proceedings of the *First Language Resources and Evaluation Conference (LREC)*. Granada, Spain.

Appendix The labels in the NE hierarchy

1 Location Names (5 subtypes)

“AST”: astronomically defined location, with physical extent
“GPL”: a (natural) geographically/geologically defined location, with physical extent
“PPL”: geo-social-political entities are politically/socially defined geographical regions
“FNC”: facility entities which are (permanent) man-made artefacts
“STR”: names of streets, avenues, roads, boulevards and postal addresses.

2 Person Names (4 subtypes)

“HUM”: human beings (alive or dead), fictional human characters etc.
“MTH”: names of saints, apostles, gods, mythical names, humanoids
“ANM”: names of animals and pets as well as mythical beasts
“CLC”: (collective) names of tribes, dynasties, ethnical and race names.

3 Organization Names (8 subtypes)

“FIN”: financial institutions, banks, capital management and funding organizations
“ATH”: org. that have an athletic dimension in their name, such as sports teams even mentions of regions in sport-related contexts
“CLT”: org. that have a cultural dimension in them, such as music, circus, theatre groups, orchestras
“PLT”: org. that have a clear political dimension in them, such as political parties, groups and movements, but also terrorist and criminal organizations and liberation armies
“TVR”: organizations that have a media profile, such as tv-channels and radio stations
“*EDU*”: educational institutions, schools, universities, academies
“*ARL*”: organizations within the air industry
“CRP”: corporations, company groups, multinational organizations, governmental organizations, non-profit organizations, governmental bodies at any level of importance, unions etc.

4 Event Names (5 subtypes)

“HPL”: historical or political, such as battles, wars, scandals, campaigns and crimes
“WTH”: events that include some kind of natural motion; weather phenomena and natural disasters such as hurricanes, cyclones, storms and typhoons
“CLU”: organized events of a cultural nature, such as festivals, conferences and fairs
“ATL”: organized events of an athletic nature, such as sports races and competitions, tournaments
“RLG”: events of a religious nature, usually a variety of holidays and special name day

5 Work And Art Names (6 subtypes)

“WRT”: names that deal with written material of type essays, studies, journals, books
“RTV”: names that denote radio and tv-programs, such as tv-series and tv-shows, radio-programs and soap operas
“WAO”: work&art names that have a physical dimension such as paintings and statues
“PRJ”: project names, agreements, initiatives
“WMD”: written media that might or might not be metonymic with the organization they represent – typical examples in this category is newspapers
“WAE”: names of operas, theater plays, symphonies

6 Object Names (8 subtypes)

“MDC”: medical and pharmaceutical products, names of drugs and medicines, but also names of diseases, proteins, genes
“FWP”: food and wine products, drinks, wines, dishes, chocolates, fruits
“CMP”: computer products both S/W and H/W as well as telephony
“VH(A|G|W)”: this sub-group (actually 3) comprises vehicles and transportation means. Depending on their primary use these are divided into water (VHW), land (VHG) or air/space vehicles (VHA)
“PRZ”: prizes (sometimes named often after people), scholarships and honours
“PRD”: general subcategory for products and artefacts, but also even names of flowers, plants etc

7 Measure Names (9 subtypes)

“VLM” volume
“TMP” temperature
“INX” index
“DST” distance
“PRC” percent
“CUR” currency
“DEN” density
“DSG” dosage
“SPD” speed

8 Time (2 subtypes)

“DAT” date
“PER” period

The UCREL Semantic Analysis System

Paul Rayson[a], Dawn Archer[b], Scott Piao[b] and Tony McEnery[b]

UCREL, Lancaster University
[a]Computing Department and [b]Department of Linguistics and Modern English Language,
Lancaster University, Lancaster, LA1 4YR
{p.rayson, d.archer, s.piao, t.mcenery}@lancaster.ac.uk

Abstract

The UCREL semantic analysis system (USAS) is a software tool for undertaking the automatic semantic analysis of English spoken and written data. This paper describes the software system, and the hierarchical semantic tag set containing 21 major discourse fields and 232 fine-grained semantic field tags. We discuss the manually constructed lexical resources on which the system relies, and the seven disambiguation methods including part-of-speech tagging, general likelihood ranking, multi-word-expression extraction, domain of discourse identification, and contextual rules. We report an evaluation of the accuracy of the system compared to a manually tagged test corpus on which the USAS software obtained a precision value of 91%. Finally, we make reference to the applications of the system in corpus linguistics, content analysis, software engineering, and electronic dictionaries.

Introduction

Understanding the meaning of words seems to present little difficulty to human beings. Indeed, children as young as seven years old seem to be able to disambiguate the various meanings of polysemous words in context. Yet, this seemingly trivial task has presented a serious challenge to the NLP research community.

Researchers in machine translation (MT) have been aware of the difficulty posed by multiple meanings of words since the 1950s and 1960s (Gale *et al*, 1993). However, whilst some researchers have allegedly left the field in frustration (Bar Hillel, for example, left when he could see no way of automatically resolving the meaning of the word *pen* in the sentence “The box was in the pen”), some others have devoted remarkable efforts to word sense disambiguation (WSD).

The WSD algorithms and systems that have been suggested and developed since the 1950s tend to draw on AI-based methods, knowledge-based methods and corpus-based methods (Ide and Véronis, 1998). However, more recently, researchers have started to combine various approaches together, as a means of obtaining better results (see, for example, Stevenson and Wilks, 2001).

A WSD system generally selects a sense from a pool of possible senses of a word that matches a given context. For example, it would tag the word “bank” as *a financial institution*[1] if it finds that the surrounding words talk about financial issues, and as *river bank* if its context talks about a river. Some WSD systems can even distinguish between “bank” as *a financial institution* and “bank” as the *building containing that institution* (or one branch of it), even though such fine-grained sense disambiguation is not always necessary within NLP (many NLP problems can be solved without access to the full set of dictionary definitions).

Let’s imagine a scenario in which we only want to know the domain of a journalistic report. In order to understand that the report talks about a crime case, it should be enough to know that many words in the news are about crime, law and the court[s]. For this type of task, what we need is a system that can determine the semantic category (or categories) of each word rather than a system that finds actual word sense definitions.

In this paper, we describe a semantic analysis system (USAS) developed at UCREL, Lancaster, which assigns semantic categories to English words. This system is different from most WSD systems in that it does not provide word meaning definitions. Rather, it assigns a semantic category to each word employing a comprehensive semantic category scheme that was originally based on the *Longman Lexicon of Contemporary English* (LLOCE) (McArthur, 1981). It is also different from the named entity identification systems, such as LaSIE in the GATE of Sheffield (Humphreys *et al*, 1999), in that it does not focus on one or two specific classes of words but, rather, assigns a tag or tags to every word in a running text. USAS combines several resources and approaches including the CLAWS POS tagger, semantic lexicons, a template list, contextual rules etc. And, as shown in our evaluation, the system performs to a high standard. Indeed, USAS obtained a precision of 91% on our evaluation corpus.

Our system has various applications in corpus linguistics and NLP. For example, it has been used to carry out content analysis of spoken and written discourse since 1990 (see Wilson and Rayson, 1993; Wilson and Leech, 1993; Wilson and Moudraia, forthcoming; Archer and Rayson, forthcoming). We have also used it to extract multiword expressions (MWE).[2] Currently, the UCREL team are incorporating USAS into an intelligent multilingual electronic dictionary, as part of the Benedict Project.[3] We believe that past experience points to wider possible applications of our system in practical NLP tasks.

[1] Definition can vary depending on the dictionary it uses.

[2] The results were extremely encouraging, particularly when extracting low-frequency MWEs (see Piao et al, 2003).

3 This is an EU project IST-2001-34237. Website: http://mot.kielikone.fi/benedict/.

Related work

The research areas closely related to our work include automatic word sense disambiguation (WSD) and semantic tagging. Research on the issue of word sense disambiguation has a long history, and a large body of literature in this area has been published. As mentioned in the previous section, approaches to WSD can generally be divided into AI-based, knowledge-based, and corpus-based ones.

The Al-based approaches were especially popular in the 1970s, but declined after the 1980s, when they were found to be impractical for large-scale language understanding (Ide and Veronis, 1998: 6-8). As large-scale lexical resources such as machine-readable dictionaries and WordNet (Fellbaum, 1998) have become increasingly available, the focus of WSD research has shifted towards WSD approaches using lexical resources (McRoy, 1992; Cowie et al, 1992; Harley and Glennon, 1997; Stevenson and Wilks, 2001).

Stevenson and Wilks (2001) provide an impressive example of a knowledge-based WSD approach. They combined several knowledge sources, tools and approaches, including LDOCE (*Longman Dictionary of Contemporary English*), a lemmatiser, a name entity identifier, Brill POS tagger, the simulated annealing optimisation algorithm (Cowie et al, 1992), selectional preferences, word subject codes and a feature extractor based on collocations and, as such, developed an "all-words" WSD system, which tags *all* content words in the input text. Stevenson and Wilks (2001) evaluated their system on the SEMCOR Corpus containing 200,000 words, and reported an accuracy of 94%.

Researchers who adopt a corpus-based approach to WSD research attempt to disambiguate word sense based on word usage information extracted from corpora (Brown et al, 1991; Yarowsky, 1995; Ng and Lee, 1996; Ng 1997). Often, statistical and machine learning algorithms are applied to distinguish different senses of a word based on pragmatic information extracted from the training corpora. Such approaches alone are unlikely to solve large-scale WSD problems. Consequently, corpus-based researchers often focus on small number of words (for example, Yarowsky (1995) conducted experiment on 12 words).

Other WSD work seeks to assign each content word with a semantic category using a pre-defined semantic taxonomy, e.g. tagging the word "father" as [HUMAN, MALE, ADULT] and "cucumber" as [NON-HUMAN, VEGETABLE], etc. A number of projects in this paradigm have been reported in the past decade, including Basili *et al.* (1997), Lowe *et al.* (1997), Lua (1997), Humphreys *et al* (1999), Demetriou and Atwell (2001).

Recently, SENSEVAL[4] has been developed to provide a framework for evaluating and comparing different WSD algorithms and systems. In spite of all these efforts, however, a generic WSD system efficient enough for practical application is yet to be developed.

[4] http://www.senseval.org/

The USAS system we present in this paper points to another generic semantic disambiguation system. Using this system, we attempt to attack the WSD problem by employing a broad semantic taxonomy rather than fine-grained word sense definitions. While such a system may fall short of orthodox WSD systems, our past experience has shown that it provides a practical means of coping with large-scale semantic disambiguation tasks. Furthermore, if we can design the same or similar semantic taxonomies for multiple languages, such a system can potentially provide a bridge for cross-language WSD and MT (cf. KAIST Multilingual WordNet (Oh et al, 2002)).

The USAS System

Architecture

Currently, the USAS system consists of the CLAWS POS tagger (Garside and Smith, 1997), a lemmatiser, a semantic tagger and some auxiliary format manipulating components. For POS tagging, we employ the C7 tagset[5]. Subsequent semantic disambiguation, to a large extent, depends on POS information encoded in this tagset. Evaluated over the large number of domains in the British National Corpus, CLAWS performs with success rates of between 96%-98%[6].

The core part of the USAS system is a semantic annotation component, which consists of semantic lexical resources, a set of context rules and programs implementing algorithms of disambiguation and assigning semantic tags to each word in a running text. The semantic lexicon resource is composed of two main parts: a single word lexicon and a collection of multi-word semantic templates. The former is used for providing candidate semantic categories for single words, while the latter is used for identifying multi-word expressions (MWE), including discontinuous MWEs, which depict single semantic concepts. Another knowledge source is a set of context rules, which provides context cues for some highly ambiguous words. Such words include "have" and "do", which can be used either as semantically significant content words or semantically "empty" function words.

USAS semantic taxonomy and tagset

The Lancaster USAS semantic tagset[7] was initially based on the *LLOCE* taxonomy, which also adopts a general ontological approach to semantic field analysis. However, it has been modified and revised in the light of practical tagging problems met in the course of applied research. This has included the splitting of several top level categories in LLOCE. For example the LLOCE top-level category "Arts and crafts, science and technology, industry and education" became three USAS top-level categories "Arts and crafts", "Science and technology" and "Education".

We have compared the scheme to other semantic category systems in detail and described the criteria underlying USAS in Archer et al (forthcoming). As USAS

[5] See http://www.comp.lancs.ac.uk/ucrel/claws7tags.html
[6] See http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2error.htm
[7] For the full tagset see http://www.comp.lancs.ac.uk/ucrel/usas/

automatically tags every word in a text, we have also added a category "Names and grammatical words" that captures words traditionally considered to be 'empty' of content (i.e. closed class words) and proper nouns. The revisions reflect our responses to problems met in light of tagging English texts from a variety of domains across different historical periods (Piao et al, 2004), and for a variety of purposes (e.g. market research, content analysis, information extraction, keyword extraction, etc.).

Currently the scheme includes 21 major discourse fields (shown in Table 1), which, in turn, expand into 232 categories. Letters are used to denote the major semantic fields while numbers are used to indicate subdivisions of the fields.

A	General & Abstract Terms
B	The Body & the Individual
C	Arts & Crafts
E	Emotional Actions, States & Processes
F	Food & Farming
G	Government & the Public Domain
H	Architecture, Building, Houses & the Home
I	Money & Commerce
K	Entertainment, Sports & Games
L	Life & Living Things
M	Movement, Location, Travel & Transport
N	Numbers & Measurement
O	Substances, Materials, Objects & Equipment
P	Education
Q	Linguistic Actions, States & Processes
S	Social Actions, States & Processes
T	Time
W	The World & Our Environment
X	Psychological Actions, States & Processes
Y	Science & Technology
Z	Names & Grammatical Words

Table 1 USAS tagset top level domains

USAS semantic lexical resources

As mentioned above, the USAS lexical resource consists of two main parts: a single word lexicon and a multi-word expression (MWE) lexicon. Currently, the former contains over 42,000 entries while the latter contains over 18,400 entries. Additionally, there is a small 'auto-tagging' single word lexicon where the entries are words containing wildcard characters. This lexicon contains around 50 entries such as '*kg' and '*km' to match weights and measures for example.

The single-word lexicon provides possible semantic categories for each word. Direct mapping between lemmas and semantic categories was not found to be viable in all cases. Stubbs (1996: 40) observed that "meaning is not constant across the inflected forms of a lemma" and Tognini-Bonelli (2001: 92) noted that lemma variants have different senses. Each word is combined with a POS tag, and they are mapped (together) to semantic categories. Since a word can have multiple POS tags in different contexts, a word may combine with each of the possible POS tags to form several entries. Fig. 2 shows some sample lexicon entries.

The MWE list aims to identify expressions such as phrasal verbs (*stubbed out*), noun phrases (*riding boots*), proper names (*United States of America*), true idioms (*living the life of Riley*) and their semantic categories. The semantic tags in template entries are arranged in the same way as in the single-word lexicon (see Fig. 3 for sample MWE lexicon entries).

```
occasion      NN1    T1.2 S1.1.1
occasion      VV0    A2.2
occasional    JJ     N6-
occasionally  RR     N6-
occult        NN1    S9
occupancy     NN1    H4
occupants     NN2    H4/S2mf M3/S2mf
occupation    NN1    I3.1 S7.1+
```

Fig 2: Sample of USAS word lexicon

```
stub*_* {Np/P*/R*} out_RP     O4.6-
ski_NN1 boot*_NN*             B5/K5.1
United_* States_N*            Z2
life_NN1 of_IO Riley_NP1      K1
```

Fig 3: Sample of USAS multiword templates

Notice that some entries are templates. These templates use simplified pattern matching codes, such as wildcards, as a means of capturing MWEs that have similar structures. For example, "*_* Ocean_N*1" will capture "Pacific Ocean", "Atlantic Ocean", etc. The templates not only match continuous MWEs, but also match discontinuous ones. In fact, numerous MWEs allow other words to be embedded within them. For example, the set phrase "turn on" may occur as "turn it on", "turn the light on", "turn the TV on" etc. Using the template " turn*_* {N*/P*/R*} on_RP " we can identify this set phrase in various contexts.

Semantic field disambiguation

As in the case of grammatical tagging, the task of semantic tagging subdivides broadly into two phases: Phase I (Tag assignment): attaching a set of potential semantic tags to each lexical unit and Phase II (Tag disambiguation): selecting the contextually appropriate semantic tag from the set provided by Phase I. USAS makes use of seven major techniques or sources of information in phase II. Below, we briefly describe the techniques (for further details, see Garside and Rayson 1997).

1. *POS tag*. Some candidate semantic tags can be eliminated by POS tagging. For example, consider the word "spring". There is a lexicon entry for spring that specifies (i) the possibility of a common noun tag, temporal noun tag or a verb tag, and (ii) the possibility that the common noun may have the 'coil' sense or the 'water source' sense. By choosing the common noun tag, the POS tagger can filter out the senses of 'jump' and 'season'. Hence the semantic tagger's task is simplified to choosing between the 'water source' and the 'coil':

word	*POS tag*	*semantic tag*
spring	temporal noun	[season]
spring	common noun	[coil] [water source]
spring	verb	[jump]

2. *General likelihood ranking for single-word and MWE tags*. The candidate senses in lexicon entries are ranked in terms of frequency, even though at present such ranking is derived from limited or unverified sources such as frequency-based dictionaries, past tagging experience and intuition. For example, "green" referring to colour is generally more frequent than "green" meaning inexperienced.

3. *Overlapping template resolution*. Normally, semantic multi-word expressions take priority over single word tagging, but in some cases a set of MWEs will produce overlapping candidate taggings for the same set of words. A set of heuristics is applied to determine the most likely MWE for tag assignment. The heuristics take account of length and span of the MWEs and how much of a template is matched in each case.

4. *Domain of discourse*. Knowledge of the current domain or topic of discourse is used to alter rank ordering of semantic tags in the lexicon and MWE list for a particular domain. Consider the adjective "battered" which has three candidate tags: 'Violence' (e.g. battered wife), 'Judgement of Appearance' (e.g. battered car), and 'Food' (e.g. battered cod). If the topic of conversation was known to be food, then we automatically raise the likelihood of the 'Food' semantic tag, at the expense of the other two tags.

5. *Text-based disambiguation*. Gale et al (1992) have used corpus analysis techniques to show that a given word largely keeps the same meaning within a text. For example, if a text uses "bank" in the sense of 'side of a river', all other occurrences of bank are likely to have that sense. In USAS, this method works together with step 4.

6. *Contextual rules*. The template mechanism is also used in identifying regular contexts in which a word is constrained to occur in a particular sense. Consider the meaning of the noun *account*: if it occurs in a sequence such as *NP's account of NP* it almost certainly means 'narrative explanation', whereas if it occurs in a financial context, in such collocations as *savings account* or the *balance of … account* it almost certainly has the meaning of a 'bank account'.

7. *Local probabilistic disambiguation*. It is generally supposed that the correct semantic tag for a given word is substantially determined by the local surrounding context. To return to the example of account: if this noun occurs in the company of words such as *financial*, *bank*, *overdrawn*, *money*, there is little doubt that the financial meaning is the correct one. However, we could identify the surrounding context not only in terms of (a) the words themselves, but also in terms of (b) their grammatical tags, (c) their semantic tags, or (d) some combination of (a) - (c). This method is still under development and future work includes experimentation, using a training corpus and a test corpus, to determine what weight to give each of these contextual factors for selecting the correct semantic tag of given word or word class. These and other factors are discussed in more detail in Garside and Rayson (1997).

Evaluation

Elsewhere, we have reported on the precision and recall of the MWE component (Piao et al, 2003), and the coverage of the lexicon across a variety of corpora (Piao et al, 2004). Here we report the breakdown of the errors for each word class and show the relative activation of the tagging methods when used in running text.

To evaluate the performance of the USAS system, we tested it on a corpus containing about 124,900 words. This corpus consists of transcriptions of 36 informal conversations, usually between two people in each case. After running the corpus though the semantic tagger, the output was manually corrected by a team of four post-editors. A team leader cross-checked post-editing decisions semi-automatically to ensure consistency within the team. Finally the machine-tagged version was compared against the hand-corrected one. Although we acknowledge that some human errors were inevitable, we assumed that human judgement is correct, and any machine outputs different from the hand-corrected version were counted as errors.

POS tag first letter	**Word class**	**Error relative to test-bed**	**Error relative to tag frequency**
A	Article	0.21	2.47
B	before clause marker	0.00	0.00
C	conjunction	0.05	0.60
D	determiner	0.21	4.69
E	existential there	0.01	1.22
F	formulae and foreign words	0.00	0.31
G	Genitive	0.01	6.62
I	preposition	0.36	4.16
J	Adjective	0.87	17.65
M	Number	0.29	23.93
N	Noun	2.62	16.29
P	Pronoun	0.06	0.51
R	Adverb	1.08	13.47
T	infinitive marker - to	0.11	7.52
U	interjection	0.02	0.94
V	Verb	3.03	13.21
X	negative	0.01	1.25
Z	Letter	0.00	2.67
Total		**8.95**	

Table 2 Breakdown of errors by POS

The rule-based methods produced a success rate of 91.05% on the post-edited test-bed. After applying the various disambiguation methods, the initial ambiguity ratio[8] of 47.73% was reduced to 17.06%. Finally, the tagger selects the first choice (most likely) tag for each word and this produces the reported error rate (8.95%). Table 2 shows the breakdown by word-class of the automatic semantic tagging errors. Such an error analysis allows us to identify where the errors occur and thus helps us to improve the accuracy of the semantic tagger.

As Table 2 illustrates, most of the errors (7.60% out of 8.95%) occurred within those word classes that relate to *content* as opposed to *function*: verb (3.03%), noun (2.62%), adverb (1.08%) and adjective (0.87%). Such a result can be expected, as the sense disambiguation of content words is generally more difficult than that of function words. The number category has the largest error rate relative to tag frequency (23.93%). This is mainly due to weights and measures being mistagged. However, because numbers occurred infrequently in our running text, they account for a mere 0.29% of the overall errors in the corpus. The tagger achieved high accuracies in respect of other word classes.

In order to examine the efficacy of the different components of the tagger, we also analysed the number of times when each component was triggered for disambiguation in running text. Table 3 shows the relative hitting rates of the 14 methods we used when tagging words and MWEs in the test corpus.

Tagging method	**Relative frequency**
Lexicon	63.68
Lexicon with stemming	3.41
Lexicon with lemmatisation	0.03
Auto-tag rule	0.39
Domain of discourse	7.67
Auxiliary verb	6.76
Context rules	0.83
Lexicon ignoring POS	0.92
Lexicon with stemming ignoring POS	0.07
WordNet unknown word look-up	0.05
Wildcard multi-word-expression	0.54
Multi-word-expression	11.60
Multi-word-expression and domain of discourse	4.06
Total	**100.00**

Table 3 Breakdown of tagging methods

Notice that, for almost 70% of the time, the semantic field was disambiguated through lexicon look-up, i.e. a combination of lexicon look-up of the surface forms and that of the stemmed or lemmatised forms. The MWE component was applied to just over 15% of words in the test corpus while the semi-automatic algorithm of assigning a domain of discourse covered almost 8%. Auxiliary verb identification appears to be particularly important since the CLAWS POS tagger does not distinguish between auxiliary and lexical verbs at the POS level. Note that, as the statistical disambiguation component is still under development, it was not included in our experiment, and hence this table does not reflect the performance of the statistical disambiguation algorithm.

[8] We define *initial ambiguity ratio* as the percentage of words in a text with more than one possible semantic tag assigned from the semantic lexicon and MWE list before the application of disambiguation techniques.

Conclusion and future work

In this paper, we described the USAS semantic tagging system. Employing a hierarchical semantic taxonomy, semantic lexical resources and a number of disambiguation algorithms such as templates, context rules etc., USAS assigns semantic categories to words and MWEs in a running text. Although different from many existing WSD systems, we believe that our system provides a practical tool for large-sale semantic annotation tasks, and that it can also support/enhance WSD systems. We also contend that such an approach would be useful for cross-language WSD and machine translation, if parallel systems were developed for other languages.

In Lancaster, further research work is under way, aiming to improve and apply the USAS system for linguistic study and language engineering tasks. For example, USAS has been used in the software engineering domain for the analysis of large volumes of technical documentation (Sawyer et al, 2002), and in decision management (Rayson et al, 2003). We are also modifying it to make it capable of tagging historical text semantically (Archer et al, 2003). Other current work includes mapping its tagset to WordNet synsets, investigating techniques to automatically detect new MWEs, and developing a mirror semantic tagger for Finnish (Lofberg et al, 2003) as part of the effort to enhance electronic dictionaries. We envisage that the USAS system will find wider applications and provide useful tool for both corpus linguistics and NLP communities.

Acknowledgements

This work is continuing to be supported by the Benedict project, EU funded IST-2001-34237. Much of the early development of the system was funded under two EPSRC projects (running between 1990 and 1996) involving Andrew Wilson and Paul Rayson and supervised by Geoffrey Leech, Roger Garside and Jenny Thomas.

References

Archer, D., Rayson, P., Piao, S., McEnery, T. (forthcoming). Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. To be presented at European Association for Lexicography 11th International Congress (Euralex 2004), Lorient, France.

Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.) Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16. (pp. 22--31) UCREL, Lancaster University.

Archer, D. and Rayson, P. (forthcoming). Using the UCREL automated semantic analysis system to investigate differing concerns in refugee literature. In proceedings of the *Keywords workshop, February 5-6, 2004*. Office for Humanities Communication, Centre for Computing in the Humanities, King's College London.

Basili, R., M. Della Rocca, and M.T. Pazienza. (1997). Towards a bootstrapping framework for corpus semantic tagging. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how?"*, Washington, D.C., April. ANLP.

Brown, P., Pietra S., Pietra, V. and Mercer, R. (1991) Word sense disambiguation using statistical methods. In Proceedings of the 29th Annual Meeting of the ACL, (pp 264-270) Berkeley, California.

Cowie, J., Guthrie, J. and Guthrie, L. (1992) Lexical Disambiguation using Simulated Annealing. *Proceedings of the 16th International Conference on Computational Linguistics (COLING-92)* (pp. 359-365) Nantes, France, July.

Demetriou, G. and Atwell, E. (2001) A domain-independent semantic tagger for the study of meaning associations in English text. In *Proceedings of the 4th International Workshop on Computational Semantics (IWCS 4)* (pp. 67-80). Tilburg, Netherlands.

Fellbaum, C. (1998) A Semantic Network of English Verbs. In Fellbaum, C. (ed.).*WordNet: An Electronic Lexical Database*. (pp. 69-104) Cambridge, Mass.: MIT Press.

Gale, W., Church, K., and Yarowsky, D. (1992), One Sense Per Discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop*. (pp.233-237).

Gale, W., Church, K. and Yarowsky, D. (1993) A method for disambiguating word senses in a large corpus. Computers and the Humanities (26), pp. 415 - 439.

Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. (pp. 102-121) Longman, London.

Garside, R. and Rayson, P. (1997) Higher-level annotation tools, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. (pp. 179-193) Longman, London.

Harley, A. and D. Glennon. (1997). Sense tagging in action: Combining different tests with additive weights. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics"*. Association for Computational Linguistics, (pp. 74-78) Washington, D.C.

Humphreys, K., Gaizauskas, R., Huyck, S., Mitchell, B., Cunningham, H., and Wilks Y. (1999) Description of the University of Sheffield LaSIE-II System as used for MUC-7**.** In Proceedings of MUC-7. Morgan Kaufmann.

Ide, N. and Veronis, J. (1998) Introduction to the special issue on word sense disambiguation: The state of art. Computational Linguistics, 24(1): 1—40.

Lofberg, L., Archer, D., Piao, S. L., Rayson, P., McEnery, T., Varantola, K., Juntunen, J-P. (2003). Porting an English semantic tagger to the Finnish language. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.) Proceedings of the CL2003 conference. UCREL technical paper number 16. (pp. 457 – 464) UCREL, Lancaster University.

Lowe, J. B. Baker, C. and Fillmore, C. (1997) 'A frame-semantic approach to semantic annotation'. In Proceedings of the SIGLEX workshop "Tagging Text with Lexical Semantics'. (pp. 18-24) Washington. D.C.

Lua. K. T. (1997) 'An efficient inductive unsupervised semantic tagger'. *Computer Processing of Oriental Languages*, 1(1), pp. 35-47.

McRoy, S. (1992). Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1-30.

McArthur, T. (1981) *Longman Lexicon of Contemporary English*. Longman, London.

Ng, H. T. and Lee, H. B. (1996) Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of AC'L96*, (pp. 40-47) Santa Cruz, CA.

Ng, H. T. (1997) Exemplar-based word sense disambiguation: some recent improvements. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP).* (pp. 208-213) Somerset, New Jersey.

Oh J-H., Saim S., Yong-Seok C.i, Key-Sun C. (2002) Word Sense Disambiguation with Information Retrieval Technique. In proceedings of LREC 2002, Las Palmas, Spain, May 2002.

Piao, S. L., Rayson, P., Archer, D., Wilson, A. and McEnery, T. (2003) Extracting Multiword Expressions with a Semantic Tagger. In *proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL 2003,* (pp. 49-56) Sapporo, Japan, July 12, 2003.

Piao, S. L., Rayson, P. Archer, D., McEnery, T. (2004). Evaluating Lexical Resources for A Semantic Tagger. *LREC 2004*, May 2004, Lisbon, Portugal.

Rayson P., Sharp B., Alderson A., et al (2003). Tracker: a framework to support reducing rework through decision management. In *Proceedings of ICEIS2003*. (pp. 344 – 351) Angers - France, April 23-26, 2003. Volume 2.

Sawyer, P., Rayson, P., and Garside, R. (2002) REVERE: support for requirements synthesis from documents. *Information Systems Frontiers Journal*. Volume 4, Issue 3, Kluwer, Netherlands, pp. 343 - 353.

Stevenson, M. and Wilks, Y. (2001) The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics* 27(3).

Stubbs, M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Blackwell, Oxford.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Benjamins, The Netherlands.

Wilson, A. and Rayson, P. (1993). Automatic content analysis of spoken discourse. In C. Souter and E. Atwell (eds.), *Corpus Based Computational Linguistics*. (pp. 215-226) Amsterdam: Rodopi.

Wilson, A. and Leech, G.N. (1993). Automatic Content Analysis and the Stylistic Analysis of Prose Literature. Revue: *Informatique et Statistique dans les Sciences Humaines 29*: 219-234.

Wilson, A. and Moudraia, O. (forthcoming) Quantitative or Qualitative Content Analysis? Experiences from a cross-cultural comparison of female students' attitudes to shoe fashions in Germany, Poland and Russia. To appear in Wilson, A., Rayson, P. and Archer, D. (eds.) *Corpus Linguistics around the world*. Rodopi, Amsterdam.

Yarowsky, D. (1995) Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of ACL-95*. (pp. 189-196) Cambridge. Massachusetts.

Towards Filling the Gap between Lexicon and Corpus in WSD

Iulia Nica♣♠, Mª Antònia Martí♣, Andrés Montoyo♦ and Sonia Vázquez♦

♣ CLiC - Centre de Llenguatge i Computació
Department of General Linguistics
University of Barcelona, Spain
iulia@clic.fil.ub.es, amarti@ub.edu
♠ Department of General Linguistics
University of Iasi, Romania

♦ Research Group of Language Processing and Information Systems
Department of Software and Computing Systems
University of Alicante, Spain
{montoyo,svazquez}@dlsi.ua.es

Abstract

Word Sense Disambiguation confronts with the lack of syntagmatic information associated to word senses: the "gap" between lexicon (here EuroWordNet, EWN) and corpus. In the present work we propose to fill this gap by applying different strategies: from one side, we extract paradigmatic information related to the ambiguous occurrence in a syntactic pattern from corpus and we incorporate it into the WSD process; from the other side, we derive discriminatory sets of senses from EWN for the ambiguous word and so we make different use of the information on senses in the lexicon; finally, we use different algorithms to map the information related to the ambiguous occurrence and the information from EWN associated to senses. Our WSD method is based on the hypothesis that meaning is principally determined by local context, thus we perform the disambiguation for the occurrences integrated into their syntactic patterns. The WSD method we propose is knowledge-driven and unsupervised. It requires only a large corpus, a minimal preprocessing phase (POS-tagging) and very little grammatical knowledge, so it can easily be adapted to other languages. We offer a synthetic overview of the research program and some preliminary tests, when applying the method on Spanish for noun disambiguation.

1. Introduction

In this article we address the issue of the existent "gap" in WSD between lexicon and corpus from the perspective of strengthening the presence of the linguistic information in the knowledge-based WSD. With this purpose, we study three variable parameters involved in the WSD process: 1) the information related to the ambiguous occurrence; 2) the information on senses contained in EWN; 3) the WSD algorithm. Our strategy to fill the gap between lexicon and corpus is to move the lexicon and the corpus one towards each other. We investigate this approximation along the following lines:

a) incorporating paradigmatic information related to the ambiguous occurrence into the WSD process and enlarging the syntagmatic information contained in the occurrence's sentence; we use the local context of the ambiguous occurrence for extracting this information from corpus;

b) making different use of the information on senses in the lexicon (EWN): we derive discriminatory sets of senses from EWN for the ambiguous word, as an alternative to the classical use of EWN for sense characterisation;

c) using different algorithms to map the information related to the ambiguous occurrence and the information from EWN associated to senses: we design an alternative algorithm that performs this mapping by exploiting the discriminatory sets for senses derived from EWN.

The combination of these parameters and of their values determines a set of WSD heuristics, as indicated above.

This kind of the investigation requires a large experimentation. We offer here a synthetic overview of this research program, with emphasis on its grounds. We also present some preliminary tests.

The WSD method we propose is a knowledge-driven and unsupervised one. It requires only a large corpus, a minimal preprocessing phase (POS-tagging) and very little grammatical knowledge, so it can easily be adapted to other languages. Up to now, we have applied the method on Spanish for noun disambiguation. The particularity of our WSD system is that sense assignment is performed using also information extracted from corpus. Thus it makes an intensive use of sense untagged corpora for the disambiguation process.

We firstly present the strategy to approximate lexicon and corpus (section 2), then the approach to WSD (section 3), the experimentation (section 4) and finally the conclusions and future work (section 5).

2. Approximating Lexicon and Corpus

2.1. From Corpus to the Lexicon

We consider that corpora contain implicit information useful for WSD. We propose a qualitative use of corpora for the extraction of information related to the word to be disambiguated, and a quantitative use to filter the obtained data. We investigate here the extraction of paradigmatic information from corpora and its use for WSD.

One way to exploit the implicit information in corpora is by means of word grouping. As a basis for clustering, we take a fundamental property of natural language: the interaction between syntagmatic and paradigmatic axes. Words that follow one another in the communicative string, oral or written, are situated on the syntagmatic axis, and establish relations that assure the coherence of the sentence. At the same time, a fixed element in a point of the syntagmatic axis can be substituted by other words, obtaining so equally coherent sentences. These virtual elements which can substitute an element in a syntagmatic string belong to a paradigmatic axis, and establish paradigmatic relations. In this way, identical syntagmatic conditions delimit word sets of paradigmatic type: the different words which can appear in a determinate position of a fixed syntagmatic pattern will have related senses, belonging to one or more common conceptual zones (Cruse, 2000: 149).

The paradigmatic relations refer to virtual elements, which could substitute a given word in the same sentence. Its identification requires a transversal look over the corpus, which goes over the limits of WSD in its usual form. Consequently, we question the word by word development for the process of WSD, dominating at present, in which the cases of ambiguity are solved independently.

We consider that disambiguation should enlarge its case by case vision to groups of occurrences or groups of words. We approximate, in this aspect, to the class-based or similarity-based methods of WSD.

In order to exploit the interaction between the syntagmatic and paradigmatic axes for a given ambiguous occurrence, it is necessary to establish the starting syntagmatic data: local context.

Context, in WSD, is usually divided in two basic categories: local and topic context. Our attention focuses on local context, on its delimitation and treatment. Local context has been exploited for sense disambiguation principally in two approaches: from a “bag of words” approach, taking into account only the lexical content words and ignoring the functional ones; or from a relational approach, using also the functional words that relate the ambiguous occurrence to the rest of lexical content units in the considered context. From this last perspective, context has been treated as n-grams (Yarowksy, 1993, Pedersen, 2001, Mihalcea, 2002) or as syntactic relations, generally limited to verb-subject and verb-object relations (Ng, 1996, Leacock *et al*., 1998, Federici *et al*., 2000, Agirre and Martínez, 2001, Martínez *et al*., 2002,), with few exceptions (Lin, 1997, Stetina *et al*., 1998).

Local context is still one of the subjects of interest in the WSD area. Some recent research focuses on issues as: the contribution of different types of information to WSD (Pedersen, 2002, Mihalcea, 2002), the use of different parameters related with context for sense tagging, and the use of algorithms for identifying the most informative parameter with respect to the sense (Hoste *et al.*, 2002, Mihalcea, 2002, Yarowsky and Florian, 2002). In spite of these studies, both the delimitation and the treatment of context were less investigated with linguistic criteria.

In our approach, local context of each ambiguous word must be set on linguistic grounds. From this perspective, we introduce the term of syntactic pattern: a triplet X-R-Y, formed by two lexical content units X and Y and an eventual relational element R, which corresponds to a syntactic relation between X and Y. Examples: [*grano*-noun *de*-preposition *azúcar*-noun], [*pasaje*-noun *subterráneo*-adjective].

Starting from the syntactic patterns of an ambiguous occurrence and looking into the corpus, we obtain different sets of words related to the occurrence. The final information we collect for the ambiguous occurrence is formed by these sets corresponding to the patterns and by the set of nouns in the sentential context. We list all the sets below:

S_1: all nouns in the sentence in which appears the ambiguous occurrence;

S_2: the whole paradigm $\{X_i\}$ corresponding to the position of the ambiguous occurrence X into a syntactic pattern P:X-R-Y; we obtain a set of type S_1 for every syntactic pattern P_k of the occurrence;

S_3: for every syntactic pattern with two nouns (that is of type [N_1-preposition-N_2], [N_1-conjunction-N_2], [N_1-comma-N_2]), the pair of these two nouns N_1 and N_2;

S_4: from every set $\{X_i\}$ of type S_1 related to a pattern P:X_0-R-Y_0, the elements that share more words $\{Y_j\}$ on the other position Y_0 of lexical content unit inside the syntactic pattern; for example, we select X_{i0} from S_1 if there is a Y_{i0} such that the patterns X_0-R-Y_{i0}, X_{i0}-R-Y_0 and X_{i0}-R-Y_{i0} do exist in the corpus; we obtain a set of type S_4 for every syntactic pattern P_k of the occurrence;

S_5: the nouns of all the syntactic patterns P_k of the ambiguous occurrence;

S_6: the intersection of the sets of type S_2;

S_7: the intersection of the sets of type S_2 and S_1;

S_8: the union of the sets of type S_2.

2.2. From Lexicon to Corpus

We are interested in make an optimal use of the paradigmatic information related to the ambiguous occurrence in the corpus. For sense assignment, we establish a mapping between this information and the paradigmatic information from the lexical source that characterises the senses. The more extensive the paradigmatic information in the lexical source, the higher the probability to perform this projection. Thus, we use a lexical source with a rich paradigmatic information, as WordNet and its multilingual variant, EuroWordNet (Vossen, 1998).

In order to potentiate the referential paradigmatic information on senses from EWN, we have developed an adaptation of the Spanish EWN, in the following way: for every sense X_i of a given word X in EWN, we extract the set SD_i of nouns related to it in EWN along the lexical-semantic relations it has. Then we eliminate the common elements (at lemma level), obtaining so disjunctive sets SD_i. As the elements of the set SD_i are related exclusively with the sense X_i, they become sense discriminators for X_i. We call the obtained lexical device "Sense Discriminators".

2.3. Mapping Corpus to the Lexicon: WSD Algorithms

Sense assignment is understood as the mapping between the information associated to the occurrence and the information provided for the word senses in EWN. Word senses in EWN are defined by their position in the net or, equivalently, by their neighbourhood: their (explicit or implicit) lexical-semantic relations with the neighbouring synsets or with the words in these synsets. From this last perspective, sense identification for a given occurrence reduces to find elements in EWN from the neighbourhood of one of its senses inside the sets related to it, previously obtained from the sentence and from corpus.

We do the mapping by means of the following WSD algorithms, corresponding to the two ways above to see a word sense in EWN:

A_1: The Specificity Mark algorithm (Montoyo and Palomar, 2000). It works on the original form of EWN. The intuitive base of this algorithm is the following: the more common information two concepts share the more related they will be. In EWN, the common information shared by two concepts corresponds to the father concept of both in the hierarchy, called Specificity Mark, SM, by the authors. The heuristic takes as input a noun set and looks for the SM in EuroWordNet with the bigger density of input words in its subtree. It chooses as correct for every input word the sense situated in the sub-tree of the SM so identified, and it lets undisambiguated the words without senses in this subtree.

A_2: The Commutative Test algorithm (Nica *et al.*, 2003). It is related to the Sense Discriminators device. At the basis of the algorithm it lays the hypothesis that if two words can commute in a given context, they have a good probability to be semantically close. From this perspective, we consider that, if an ambiguous occurrence can be substituted in a syntactic pattern by a sense discriminator, then it can have the sense corresponding to that sense discriminator. We call this algorithm the Commutative Test (CT). In order to reduce the computational cost of this substitution operation, we perform an equivalent process: We previously extract, from corpus, the possible substitutes of the ambiguous occurrence in a syntagmatic pattern, and then we intersect this set with every set of sense discriminators; the senses for which the intersection is not empty can be assigned to the occurrence. When applied on a set S of words, the algorithm intersects it with every set SD_i; if it obtains a not empty intersection between S and SD_{i0}, then it concludes that X can have the sense X_{i0} in the starting syntactic pattern.

These two algorithms make a different use of EWN: SM exploits the hierarchy and only the hipo/hiperonymy relations, meanwhile CT is equivalent to rather a radial perspective on EWN and it also operates with mero/holonymy, synonymy and co-hyponymy (or coordination) relations.

3. Strategy for WSD

The previous considerations lead us to a different approach to WSD: the occurrence to be disambiguated is considered not separately, but integrated into a syntactic pattern, and its disambiguation is carried out in relation to this pattern. In this approach, the integration of a word occurrence into local syntactic patterns is a first approximation to its meaning in context.

Our strategy is based on the hypothesis that the local syntactic patterns in which an ambiguous occurrence participates have decisive influence on its meaning and thus they are highly relevant for the sense identification. We assume that inside a syntactic pattern a word will tend to have the same sense: the "quasi one sense per syntactic pattern" hypothesis.

The integration of the ambiguous occurrence in a local syntactic pattern constitutes the key element of our proposal for bringing together the paradigmatic information in the lexicon and the syntagmatic information identifiable in the context. On the grounds of the syntactic patterns, we identify in the corpus the set of the possibilities for the position of the ambiguous occurrence into the syntactic pattern, obtaining so a word class of paradigmatic type. We can thus incorporate paradigmatic information into the WSD process together with the traditional syntagmatic information, for a better mapping between corpus and lexicon: we apply on the class previously obtained a disambiguation algorithm based on paradigmatic relations from EWN.

The method works as follows:

1) the identification of the syntactic patterns of the ambiguous occurrence;

2) the extraction of information related to it: from corpus and from the sentential context;

3) the application of the WSD algorithms (SM, CT) on the information previously obtained.

We detail these steps in the next subsections.

3.1. Syntactic Patterns Identification

In order to identify and exploit the syntactic patterns of an ambiguous occurrence, we first define a list of basic patterns, in terms of parts-of-speech (POS), that covers a subset of the possible syntactic relations involving nouns. The identification of the syntactic patterns for an occurrence its done following two criteria: a) a structural one (we considering only the sequences corresponding to one of the predefined morphosyntactic patterns) and b) of frequency (we keep only those sequences which appear more times in the search corpus).

3.2. Extraction of Information Related to the Ambiguous Cccurrence

We detail here only the modality to extract the sets S_2 and S_4 of the paradigm associated to a position of a lexical content element into a syntactic pattern; the extraction of the other sets S_i is trivial, either from the sentence (sets S_1) either from sets S_2 and S_4.

S_2: The paradigm is obtained by fixing the syntactic pattern at lemma and morphosyntactic levels, and letting variable only the position of the ambiguous word at lemma level.

S_4: For the ambiguous occurrence into a syntactic pattern P: X-R-Y, we first fix the X position and let variable the other position Y for the search into the corpus; then we fix the Y position for the every one of the variants previously found for it and let variable the X position for the search into the corpus. Inside the set of possible substitutes of X, including X, we delimit groups of nouns that share Y and some other(s) substitute(s) of Y on the other position inside the syntactic pattern.

3.3. Sense Assignment

We have developed several WSD heuristics; they are determined by the combination between a set S_i (section 2.1.) and an algorithm A_j (section 2.3). Thus the WSD system is a set of heuristics $H_{ij} = (S_i, A_j)$.

In order to obtain a high reliability in the sense assignment, we construct a complex WSD system that incorporates the designed heuristics as voters. In every one of the heuristics we keep all the proposed sense for the ambiguous occurrence, and the final sense assignment is established on the basis of the number of voters for every individual sense.

3.4. Example

We illustrate the method for noun *órgano* in the occurrence number 35 from Senseval-2:

Los enormes y continuados progresos científicos y técnicos de la Medicina actual han logrado hacer descender espectacularmente la mortalidad infantil, erradicar multitud de enfermedades hasta hace poco mortales, sustituir mediante trasplante o implantación <head>órganos</head> dañados o partes del cuerpo inutilizadas y alargar las expectativas de vida.

The steps of the disambiguation process are the following:

0. Preprocessing

a. Input text POS-tagging

b. Extraction of Sense Discriminators sets

In EWN, *órgano* has five senses[1]:

órgano_1: 'part of a plant';

órgano_2: 'governmental agency, instrument';

órgano_3: 'functional part of an animal';

órgano_4: 'musical instrument'

órgano_5: 'newspaper'.

Correspondingly, we obtain from the EWN hierarchy the following Sense Discriminators sets:

SD1: {*órgano vegetal, espora, flor, pera, manzana, bellota, hinojo, semilla, poro, …*}

SD2: {*agencia, unidad administrativa, banco central, servicio secreto, seguridad social, …*}

SD_3: {*parte del cuerpo, trozo, músculo, riñón, oreja, ojo, glándula, lóbulo, tórax, dedo, …*}

SD_4: {*instrumento de viento, instrumento musical, mecanismo, teclado, pedal, …*}

SD_5: {*periódico, publicación, medio de comunicación, serie, número, ejemplar, …*}

1º. Syntactic patterns identification for the ambiguous occurrence:

Using search schemes and decomposition rules associated to the syntactic patterns, we find the sequence [*órganos*-N *dañados*-ADJ *o*-CONJ *partes*-N] and from this we extract two basic patterns: [*órgano*-N *o*-CONJ *parte*-N] and [*órgano*-N *dañado*-ADJ].

2º. Extraction of information associated to the ambiguous occurrence:

2a. From the context, we extract the nouns of the sentence:

S_1 = {*progreso, científico, mortalidad, multitud, enfermedad, mortal, trasplante, implantación, órgano, parte, cuerpo, expectativa, vida*}

2b. From corpus, we extract the paradigm corresponding to the position of *órgano* in each of the two syntactic patterns previously identified. To do this, we let vary, at lemma level, the position of *órgano* in the two patterns: [*X*-N *o*-CONJ *parte*-N] and [*X*-N *dañado*-ADJ] respectively. With the help of the search schemes, we then look in the corpus for the possible nouns as X in any of the possible realisations of these two patterns. We obtain two sets whose reunion is the following:

S_2 = {*mediador, terreno, chófer, árbol, cabeza, planeta, parte, incremento, totalidad, guerrilla, programa, mitad, país, temporada, artículo, tercio*}

3º WSD algorithms application. WSD heuristics

3.1. Specificity Mark:

H_{11}: By applying the SM algorithm on S1, we obtain sense 4 in EWN for *órgano*, which corresponds to sense 1 in Senseval.

3.2. Commutative Test application:

[1] The pseudodefinitions are ours.

H_{12}: $S_1 \cap SD_1 = \varnothing$; $S_1 \cap SD_2 = \varnothing$; $S_1 \cap SD_3 \neq \varnothing$; $S_1 \cap SD_4 = \varnothing$; $S_1 \cap SD_5 = \varnothing$. Thus, the heuristics concludes that X can have sense 3.
H_{82}: $S_8 \cap SD_1 = \varnothing$; $S_8 \cap SD_2 = \varnothing$; $S_8 \cap SD_3 \neq \varnothing$; $S_8 \cap SD_4 = \varnothing$; $S_8 \cap SD_5 = \varnothing$. Thus, the heuristics also concludes that X can have sense 3.
4º. Final sense assignment:
In this case, we obtain the same sense 3 from both heuristics, so we assign sense 3 from EWN to the occurrence of *órgano*, which corresponds to the correct sense 2 in the Senseval-2 dictionary.

4. Experimentation and discussion

We have planned a series of experiments, with focus on various aspects of our proposal for WSD: a) the contribution of each type of information associated to the ambiguous occurrence to the disambiguation process; b) the efficiency of each of the algorithms and their eventual complementarity in the use of information in EWN as well as of the information associated to the occurrence; c) useful combinations of WSD algorithms and of word sets associated to the occurrence; d) optimal variants to unify different heuristics into a complex WSD system. In parallel, we analyse the results from a linguistic perspective, for a better understanding of the impact on the DSA process from the elements there involved and from their interrelation. We are also interested into the implications that our method can have on the issue of the sense characterisation and delimitation for WSD.
The projected experiments are under development. At the present, we have performed a few initial tests that we selectively present below. Our principal purpose in these initial tests has been to test the usefulness of the paradigmatic information for the WSD process vs. the traditional syntagmatic information as well as the efficiency of the Sense Discriminators device and the associated algorithm, the Commutative Test. Thus we have limited until now to work with two control word sets associated to an ambiguous occurrence: for the syntagmatic information, we have used S_1; for the paradigmatic information we have used an alternative set, S_8, the union of the sets of type S_2 obtained for all the syntactic patterns. We have performed all the experiments on the test corpus from the Spanish Senseval-2 exercise, for an objective evaluation.
We have first worked with only very few basic syntactic patterns and search schemes, and used, as search corpus, LEXESP (5,5 millions words). The application of the SM algorithm on set S_8 led to the results in table 1:

A_1: Specificity Mark	Precision	Recall	Coverage
S_8	45,7%	7,5%	16,4%

Table 1: Results of Specificity Marks Algorithm

Although we have obtained low results with respect to the level reached in the Senseval-2 exercise (precision of 51,4%-71,2%, recall of 50,3%-71,2%, coverage of 98%-100%), the test demonstrates that the paradigmatic information is really useful for WSD.
In a second group of experiments, we have verified the CT algorithm. We have also tried to improve some parameters of the method: we have adopted a larger search corpus (EFE, over 70 millions words) and progressively enlarged the basic syntactic patterns and the search schemes. The results were those in table 2:

A_2: Commutative Test	Precision	Recall	Coverage
S_8	54,1%	11,6%	21,4%
S_1	59,6%	4,7%	7,9%
$S_8 + S_1$	56,1%	15,2%	27,1%

Table 2: Results of Commutative Test Algorithm

This group of tests has shown, first, that we do can make WSD by means of the Sense Discriminators device and the Commutative Test. Secondly, the use of the paradigmatic information sensibly improves the performance of the WSD algorithm. We have thus a confirmation of our strategy to incorporate paradigmatic information from corpus in the WSD process. In terms of precision, the performance of the two heuristics is practically the same. This suggests that the two types of information, paradigmatic (S_8) and syntagmatic (S_1), are equally useful for sense assignment, thus it is necessary to exploit them both in WSD tasks. Thirdly, the incremented size of the corpus leads to better results, in all three evaluation criteria: precision, recall and coverage.
We analysed the method, step by step; we present below some of the observations.
1) The level of disambiguation is highly affected both by the quantity and the quality of the syntactic patterns identification. In this experiment, we have identified syntactic patterns for only 70% of the occurrences to be disambiguated, as we have considered only a part of the possible structural patterns. We haven't used any qualitative filter on these patterns, and thus we have obtained coverage with answers for only a 29% of them. For this reason, we do believe that the real potential of our method is higher and so the improvement of patterns delimitation is a stringent necessity.
2) As we have designed, but not yet tested, a filter for the syntactic patterns based on their frequency in the search corpus, we have analysed the quality of the repeated patterns in the Senseval-2 test corpus. Indeed, they do are correct patterns. This demonstrates that the frequency is a useful criterion to take into consideration in order to improve the patterns identification.
3) We have also verified, for the iterative syntactic patterns in the Senseval-2 test corpus, the "quasi one sense per syntactic pattern" hypothesis. The data is very

limited, as we have identified for the moment only 45 identified iterative syntactic patterns, but it seems that there is a tendency of the syntactic patterns to associate with a unique word sense (44 cases on 45). This gives some preliminary support to our strategy of integrating the ambiguous occurrences into syntactic patterns as a first step towards their disambiguation.

5. Conclusions and Future Work

In this paper we propose an approach to WSD that overcomes the gap between lexicon and corpus that affects the knowledge-based methods. We do it, by means of the intensive use of sense-untagged corpora following linguistic criteria. The characteristics of the approach are:
- independence of a corpus annotated at the sense level;
- expansive disambiguation: of any occurrence of a given word in the same syntactic pattern, and of the substitutes of the word in the pattern;
- partial reduction of data sparseness problem: as there are considered the different syntactic patterns in which an ambiguous occurrence appears, there are more probabilities to obtain, from corpus, information related to the occurrence, for each one of the patterns;
- transferability from one language to another with minimal costs;
- linguistic grounds: there are exploited properties of language.

The future experiments we have designed are orientated towards:
- coverage improvement;
- quality improvement (filters on the patterns and on the extracted paradigms);
- identification of optimal modalities to select the final sense assignment on the basis of multiple heuristics;
- combination with some other type of information, principally related to the domain;
- interpretation of the linguistic implications.

6. Bibliography

• Agirre, E. and D. Martínez, 2001. Learning class-to-class selectional preferences. In: *Proceedings of the ACL CONLL'2001 Workshop*, Tolouse

• Civit, M., 2003. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*, Ph.D. Thesis, University of Barcelona

• Corazzari, O., N. Calzolari, and A. Zampolli, 2000. An Experiment of Lexical-Semantic Tagging of an Italian Corpus. In: Proceeding of *LREC'2000*, Athens

• Cruse, Alan, 2000. *Meaning in Language. An Introduction to Semantics and Pragmatics*, Oxford University Press

• Hoste, V., I. Hendrickx, W. Daelemans and A. van den Bosch, 2002. Parameter optimisation for machine-learning of WSD. In: *Natural Language Engineering*, ***8(4)***

• Kilgariff, A., 1998. Bridging the gap between lexicon and corpus: convergence of formalisms. In: *Proceedings of LREC'1998*, Granada

• Leacock, C., M. Chodorow and G.A. Miller, 1998. Using Corpus Statistics and WordNet Relations for Sense Identification, *Computational Linguistics. Special Issue on Word Sense Disambiguation*, **24 (1)**

• Lin, D., 1997. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In: *Proceedings of ACL and EACL'97*, San Francisco

• Manning, C. and H. Schütze, 1999. *Foundations of Statistical Natural Language Processing*, MIT Press

• Martínez D., E. Agirre E. and L. Màrquez L, 2002. Syntactic Features for High Precision Word Sense Disambiguation. In: *Proceedings of COLING'02*, Taipei

• Mihalcea, R. and D. Moldovan, 1999. An Automatic Method for Generating Sense Tagged Corpora. In: *Proceedings of AAAI '99*, Orlando

• Montemagni, S., S. Federici and V. Pirelli, 1996. Example-based Word Sense Disambiguation: a Paradigm-driven Approach. In: *Proceedings of EURALEX'96*, Göteborg

• Montoyo, A. and M. Palomar, 2000. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. In: *Proceedings of DEXA'00*, Greenwich

• Ng, H.T. and H.B. Lee, 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In: *Proceedings ACL'96*, Santa Cruz

• Nica, I., M. A. Martí and A. Montoyo, 2003. Colaboración entre información paradigmática y sintagmática en la Desambiguación Semántica Automática, *XX Congreso de la SEPLN 2003*, Alcalá de Henares. In: *Revista de Procesamiento del Lenguaje Natural*, **31**

• Pedersen, T., 2001. A decision tree of bigrams is an accurate predictor of word sense. In: *Proceedings of NAACL 2001*, Pittsburg

• Pedersen, T., 2002. A Baseline Methodology for Word Sense Disambiguation. In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, February, Mexico City

• Stetina, J., S. Kurohashi and M. Nagao, 1998. General WSD Method based on a Full Sentential Context. In: *Proceedings of COLING-ACL Workshop*, Montreal

• Véronis, J., 2001. *Sense tagging: does it make sense?*. Paper presented at the Corpus Linguistics'2001 Conference, Lancaster

• Vossen, P., 1998 (ed.). *EUROWORDNET. A Multilingual Database with Lexical Semántic Networks*, Kluwer APademic Publishers, Dordrecht

• Yarowsky, D., 1993. One Sense per Collocation. In: *DARPA Workshop on Human Language Technology*, Princeton

• Yarowsky, D. and R. Florian, 2002. Evaluating sense disambiguation across diverse parameter spaces. In: *Natural Language Engineering,* ***8(4),*** Cambridge University Press

Semantic Annotation for Interlingual Representation of Multilingual Texts

Teruko Mitamura[1], Keith Miller[2], Bonnie Dorr[3], David Farwell[4], Nizar Habash[3], Stephen Helmreich[4], Eduard Hovy[5], Lori Levin[1], Owen Rambow[6], Florence Reeder[2], Advaith Siddharthan[6]

[1.] Carnegie Mellon University {teruko,lsl}@cs.cmu.edu, [2.] MITRE Corporation {keith,freeder}@mitre.org, [3.] University of Maryland {bonnie,nizar}@umiacs.umd.edu, [4.] New Mexico State University {david,shelmrei}@crl.nmsu.edu, [5.] University of Southern California, <hovy@isi.edu>, [6.] Columbia University {rambow, as372}@cs.columbia.edu

Abstract

This paper describes the annotation process being used in a multi-site project to create six sizable bilingual parallel corpora annotated with a consistent interlingua representation. After presenting the background and objectives of the effort, we describe the multilingual corpora and the three stages of interlingual representation being developed. We then focus on the annotation process itself, including an interface environment that supports the annotation task, and the methodology for evaluating the interlingua representation. Finally, we discuss some issues encountered during the annotation tasks. The resulting annotated multilingual corpora will be useful for a wide range of natural language processing research tasks, including machine translation, question answering, text summarization, and information extraction.

1 Introduction

An interlingua is a semantic representation which mediates between source and target languages in interlingua-based machine translation. It is designed to capture the meaning of a sentence that is common to both source and target languages. If a system supports multi-language translation, the design of the interlingua becomes more complex, due to the number of languages represented. Even though the aim of an interlingua is to capture language-independent semantic expressions, it is difficult to design an interlingua that covers all known languages, and there is no universally acceptable interlingua representation currently in existence. In practice, researchers have designed interlingua representations for particular sets of languages, in order to cover the necessary set of semantic expressions for machine translation (Mitamura et al. 1991). More recently, the use of interlingua representations has been extended beyond machine translation to include, for example, applications for question answering (Ogden et al., 1999), representing agent actions (Kipper & Palmer, 2000) and knowledge acquisition from text (Nyberg et al. 2002).

In September 2003, researchers from six sites began a project titled "Interlingual Annotation of Multilingual Corpora" (IAMTC) [1], funded by the National Science Foundation. This project focuses on the creation of a semantic representation system, followed by the development of six semantically-annotated bilingual corpora. The bilingual corpora pair English texts with corresponding text in Japanese, Spanish, Arabic, Hindi, French, and Korean. The semantically annotated corpora will be useful not only for machine translation development, but also for research in question answering, text summarization and information retrieval. The project participants include the Computing Research Laboratory at NMSU, the Language Technologies Institute at CMU, the Information Sciences Institute at USC, UMIACS at the University of Maryland, the MITRE Corporation, and Columbia University.

In this paper, we first present the objectives of the IAMTC project. We then provide background information on the multilingual corpora and the three stages of interlingual representation being developed. We then focus on the annotation process itself, including a description of an interface environment that supports the annotation task, and a discussion of the evaluation methodology. We conclude with a summary of the current status of the project, and discuss some issues encountered during the annotation tasks.

2 Project Goals

The IAMTC project has the following goals:

- Development of an interlingua representation framework based on a careful study of text corpora in six languages and their translations into English.
- Development of a methodology for accurately and consistently assigning such representations to texts across languages and across annotators.
- Annotation of a corpus of six multilingual parallel subcorpora, using the agreed-upon interlingual representation.
- Development of semantic annotation tools which serve to facilitate more rapid annotation of texts.
- Design of new metrics and evaluations for the interlingual representations, in order to evaluate the

[1] http://aitc.aitcnet.org/nsf/iamtc/

degree of annotator agreement and the granularity of meaning representation.

3 Corpus

The data set consists of 6 bilingual parallel corpora. Each corpus is made up of 125 source language news articles along with three independently produced translations into English. (The source news articles for each individual language corpus are different from the source articles in the other language corpora.) The source languages are Japanese, Korean, Hindi, Arabic, French and Spanish. Typically, each article contains between 300 and 400 words (or the equivalent) and thus each corpus has between 150,000 and 200,000 words. Consequently, the size of the entire data set is around 1,000,000 words. The Spanish, French, and Japanese corpora are based on the DARPA MT evaluation data (White and O'Connell 1994). The Arabic corpus is based on LDC's Multiple Translation Arabic, Part 1 (Walker et al., 2003).

For any given subcorpus, the annotation effort is to assign interlingual content to a set of 4 parallel texts (one in the original source language, plus 3 translations to English by different translators), all of which theoretically communicate the same information. A multilingual parallel data set of source language texts and English translations offers a unique perspective and unique problem for annotating texts for meaning.

4 Interlingua

The interlingual representation comprises three levels and incorporates knowledge sources such as the Omega ontology (Philpot et al., 2003) and theta grids (Dorr, 2001). The three levels of representation are referred to as *IL0, IL1* and *IL2*. The aim is to perform the annotation process incrementally, with each level of representation incorporating additional semantic features and removing existing syntactic ones. *IL2* is intended as the interlingual level that abstracts away from (most) syntactic idiosyncrasies of the source language. *IL0* and *IL1* are intermediate representations that are useful stepping stones for annotating at the next level.

4.1 IL0

IL0 is a deep syntactic dependency representation. It includes part-of-speech tags for words and a parse tree that makes explicit the syntactic predicate-argument structure of verbs. The parse tree contains labels referring to deep-syntactic grammatical function (normalized for voice alternations). *IL0* does not contain function words (their contribution is represented as features) or semantically void punctuation. While this representation is purely syntactic, many disambiguation decisions, relative clause and PP attachment for example, have been made, and the presentation abstracts as much as possible from surface-syntactic phenomena. (Thus, our *IL0* is intermediate between the analytical and tectogrammatical levels of the Prague School (Hajič et al 2001).) IL0 is constructed by hand-correcting the output of a dependency parser (see section 6), and allows annotators to see how textual units relate syntactically when making semantic judgments. Thus, it is a useful starting point for semantic annotation at *IL1*.

4.2 IL1

IL1 is an intermediate semantic representation. It associates semantic concepts with lexical units like nouns, adjectives, adverbs and verbs. It also replaces the syntactic relations in *IL0*, like *subject* and *object*, with thematic roles, like *agent, theme* and *goal*. Thus, like PropBank (Kingsbury et al 2002), *IL1* neutralizes different alternations for argument realization. However, *IL1* is not an interlingua; it does not normalize over all linguistic realizations of the same semantics. In particular, it does not address how the meanings of individual lexical units combine to form the meaning of a phrase or clause. It also does not address idioms, metaphors and other non-literal uses of language. Further, *IL1* does not assign semantic features to prepositions; these continue to be encoded as syntactic features of their objects, which may be annotated with thematic roles such as *location* or *time.*

4.3 IL2

IL2 is intended to be an interlingua, a representation of meaning that is (reasonably) independent of language. IL2 is intended to capture similarities in meaning across languages and across different lexical/syntactic realizations within a language. For example, like FrameNet (Baker et al 1998), IL2 is expected to normalize over conversives (e.g. X bought a book from Y vs. Y sold a book to X) and also over non-literal language usage (e.g. X started its business vs. X opened its doors to customers). The exact definition of IL2 is the major research contribution of this project. However, it is important to note that even at the level of IL2, it does not include more complex linguistics phenomena, such as speech acts, discourse analysis and pragmatics.

4.4 The Omega Ontology

In progressing from IL0 to IL1, annotators select semantic terms (concepts) to represent the nouns, verbs, adjectives, and adverbs present in each sentence. These terms are represented in the 110,000-node Omega ontology (Philpot et al., 2003), under construction at ISI. Omega has been built semi-automatically from a variety of sources, including Princeton's WordNet (Fellbaum, 1998), New Mexico State University's Mikrokosmos (Mahesh and Nirenburg, 1995), ISI's Upper Model (Bateman et al., 1989) and ISI's SENSUS (Knight and Luk, 1994). The ontology, which has been used in several projects in recent years (Hovy et al., 2001), can be browsed using the DINO browser at http://blombos.isi.edu:8000/dino; this browser forms a part of the annotation environment. Omega continues to be developed and extended.

4.5 The Theta Grids

Each verb in Omega is assigned one or more theta grids specifying the theta roles of arguments associated with that verb. Theta roles are abstractions of deep semantic relations that generalize over verb classes. They are by far the most common approach in the field to represent predicate-argument structure. However, there are numerous variant theories with little agreement even on terminology (Fillmore, 1968; Stowell, 1981; Jackendoff, 1972; Levin and Rappaport-Hovav, 1998).

The theta grids used in our project were extracted from the Lexical Conceptual Structure Verb Database (LVD) (Dorr, 2001). The "LCS Database" contains Lexical conceptual Structures built by hand, organized into semantic classes that are a reformulated version of those in Beth Levin (1993) English Verb Classes and Alternations (EVCA), Part 2. The WordNet senses assigned to each entry in the LVD link the theta grids to the verbs in the Omega ontology. In addition to the theta roles, the theta grids specify syntactic realization information, such as Subject, Object or Prepositional Phrase, and the Obligatory/Optional nature of the argument. The set of theta roles used, although based on research in LCS-based MT (Dorr, 1993; Habash et al, 2002), has been simplified for this project.

5 Annotation Tools

We have assembled a suite of tools to be used in the annotation process. Since we are gathering our corpora from disparate sources, we need to standardize the text before presenting it to automated procedures. For English, this involves sentence boundary detection, but for other languages, it may involve segmentation, chunking of text, or other operations. The text is then processed with a dependency parser, the output of which is viewed and corrected in TrED (Hajič, et al., 2001), a graphically-based tree editing program, written in Perl/Tk2. The revised deep dependency structure produced by this process is the IL0 representation for that sentence.

To create IL1 from the IL0 representation, annotators use Tiamat, a tool developed specifically for this project. This tool enables viewing of the IL0 tree with easy reference to all of the IL resources described in section 4 (current IL representation, ontology, and theta grids). Tiamat provides the ability to annotate text via simple point-and-click selections of words, concepts, and theta-roles. The IL0 is displayed in the top left pane, ontological concepts and their associated theta grids, if applicable, are located in the top right, and the sentence itself is located in the bottom right pane. An annotator may select a lexical item (leaf node) to be annotated in the sentence view; this word is highlighted, and the relevant portion of the Omega ontology is displayed in the pane on the left. In addition, if this word has dependents, they are automatically underlined in red in the sentence view. Annotators can view all information pertinent to the process of deciding on appropriate ontological concepts in this view. Following the procedures described in section 6, selection of concepts, theta grids and roles appropriate to that lexical item can then be made in the appropriate panes.

In order to evaluate the annotators' output, an evaluation tool was also developed to compare the output and to generate the evaluation measures that are described in section 7. The reports generated by the evaluation tool allow the researchers to look at both gross-level phenomena, such as inter-annotator agreement, and at more detailed points of interest, such as lexical items on which agreement was particularly low, possibly indicating gaps or other inconsistencies in the ontology.

[2] http://quest.ms.mff.cuni.cz/pdt/Tools/Tree_Editors/Tred/

6 Annotation Manuals and Process

To describe the annotation task, we first present the annotation manuals and then discuss the annotation process.

6.1 Annotation Manual

We have been developing markup instructions which comprise three manuals: a users' guide for Tiamat (including procedural instructions), a definitional guide to semantic roles, and a manual for creating a dependency structure (IL0). Together these manuals allow the annotator to understand (1) the intention behind aspects of the dependency structure; (2) how to use Tiamat to mark up texts; and (3) how to determine appropriate semantic roles and ontological concepts. In choosing a set of appropriate ontological concepts, annotators were encouraged to look at the name of the concept and its definition, the name and definition of the parent node, example sentences, lexical synonyms attached to the same node, and sub- and super-classes of the node.

6.2 Annotation process

For the initial testing period, only English texts were annotated, and the process described here is for English text. We assume that the process for non-English texts would be the same with a minor modification as needed.

Each sentence of the text is parsed into a dependency tree structure. For English texts, these trees were first provided by the Connexor parser (Tapanainen and Jarvinen, 1997), and then corrected by one of the team PIs. Then the corrected dependency structures (IL0) are provided to annotators.

The annotators were instructed to annotate all nouns, verbs, adjectives, and adverbs. This involves choosing all relevant concepts from Omega – both concepts from Wordnet SYNSETs and those from Mikrokosmos; these sources of information are intertwined in Omega. One of the goals and results of this annotation process will be a simultaneous coding of concepts in both ontologies, facilitating a closer union between them.

In addition, annotators were instructed to provide a semantic case role for each dependent of a verb. LCS verbs were identified with Wordnet classes and the LCS case frames were supplied where possible. The annotator, however, was often required to determine the set of roles or alter them to suit the text. In both cases, the revised or new set of case roles was noted and sent to a reviewer for evaluation and possible permanent inclusion. Thus the set of event concepts in the ontology supplied with roles will grow through the course of the project.

For the initial testing phase of the project, all annotators at all sites worked on the same texts. We have two annotators from each site. Each site, which has different source language texts, provided two texts that were translated into English by two different translators. To test for the effects of coding two texts that are semantically close (since they are both translations of the same source document), the order in which the texts were annotated differed from site to site. Half of the sites marked one translation first, and the other half of the sites marked the second translation first. Another variant tested was to interleave the two translations, so that two similar sentences were coded consecutively.

In the period leading up to the initial test phase, weekly conversations were held at each site by the annotators to review the coded texts. This was followed by a weekly conference call among all the annotators. During the test phase, no discussion was permitted until all the annotation tasks were completed.

7 Evaluation Methodology

We have identified several metrics for evaluation of intercoder agreement on annotations. We are currently measuring intercoder agreement on concept names selected from the Omega ontology and thematic role labels.

Two measures of intercoder agreement are currently used, Kappa (Carletta, 1993) and a Wood Standard similarity (Habash and Dorr, 2002). For expected agreement in the Kappa statistic, P(E) is defined as 1/(N+1) where N is the number of choices at a given data point. In the case of Omega nodes, this means the number of matched Omega nodes (by string match) plus one for the possibility of the annotator traversing up or down the hierarchy. The Wood Standard is the category chosen by the most annotators. In cases of no agreement, a random selection is picked from the annotator's selections. Multiple measures were used because it is important to have a mechanism for evaluating inter-coder consistency in the use of the IL representation language which does not depend on the assumption that there is a single correct annotation of a given text.

In addition to intercoder agreement, we are also developing metrics for evaluating the quality of an annotated interlingua. Given the project goal of generating an IL representation which is useful for MT (among other NLP tasks), we measure the ability to generate accurate surface texts from the IL representation as annotated. At this stage, we plan to use an available generator, Halogen (Knight and Langkilde, 2000). A tool to convert the representation to meet Halogen's requirements is being built. Following the conversion, surface forms will be generated and then compared with the originals through a variety of standard MT metrics (ISLE, 2003). This will serve to determine whether the elements of the representation language are sufficiently well-defined and whether they can serve as a basis for inferring interpretations from semantic representations or (target) semantic representations from interpretations.

8 Annotation Issues

During the test phase, we annotated 144 texts, which come from 2 translations of 6 source texts annotated by 2 annotators in each 6 sites.

A preliminary investigation of intercoder agreement on multiple annotations shows that the more annotators learn the process, the better they become, resulting in an improvement of intercoder agreement. We made two assumptions regarding the training of novice annotators in order to improve intercoder agreement. One assumption is that novice annotators may make inconsistent annotations within the same text. In order to train annotators, we have developed an intra-annotator consistency checking procedure. After the annotators finished an initial annotation pass, they were asked to go over their results to see if there were any inconsistencies within the text. For example, if two nodes in different sentences are co-indexed, then annotators must ensure that the two nodes carry the same meaning in the context of the two different sentences.

Another assumption we made was that if two annotators at the same site discuss their annotation results after their annotation tasks are completed, they can learn more from each other. Under this assumption, we have developed inter-annotator a reconciliation procedure and a voting tool associated with this process. There are three steps to follow. First, we created a combined annotation file, in which disagreements are marked in red. Each annotator votes privately either Yes, Possible, or No for items marked in red. In the second step, annotators get together and discuss the differences. After the open discussion, they vote again privately. We are currently in the process of analyzing the effect of inconsistency checking and inter-annotator reconciliation on overall intercoder agreement.

During the inter-annotator reconciliation process, we have encountered a number of difficult issues. One issue is the granularity of concept selection. The Omega ontology, which is derived from WordNet, contains 110,000 nodes and often provides too many alternatives, whereas Omega-Mikrokosmos, which contains only 6,000 concepts, does not offer all the concepts needed for annotation. For example, the word extremely contains 4 concepts in Omega's WordNet, and each of the senses is hard to distinguish from the others: (1) to a high degree or extent; favorably or with much respect, (2) to an extreme degree, (3) to an extreme degree, super, (4) to an extreme degree or extent, exceedingly. On the other hand, Omega-Mikrokosmos does not contain a concept for the word extremely.

In the coming months we will be pruning out the extraneous terms from Omega, fleshing out the current procedures for evaluating the accuracy of an annotation and measuring the inter-coder agreement. We will also be working on IL2 design and annotation. Finally, a growing corpus of annotated texts at each stage (IL0, IL1, IL2) will become available.

Additional issues to be addressed include: (1) personal name, temporal and spatial annotation (e.g., Ferro et al., 2001); (2) causality, co-reference, aspectual content, modality, speech acts, etc; (3) reducing vagueness and redundancy in the annotation language; (4) inter-event relations such as entity reference, time reference, place reference, causal relationships, associative relationships, etc; Finally, to incorporate these, cross-sentence phenomena remain a challenge.

From an MT perspective, issues include evaluating consistency in the use of an annotation language, given that any source text can result in multiple, different, legitimate translations (Farwell and Helmreich, 2003). Along these lines, there is the problem of annotating texts for translation without including in the annotations inferences from the source text.

8 Conclusion

The IAMTC project is radically different from those annotation projects that have focused on morphology, syntax or even certain types of semantic content (e.g., for word sense disambiguation). It is most similar to PropBank (Kingsbury et al 2002) and FrameNet (Baker et al 1998). However, our project is novel in its emphasis on: (1) a more abstract level of mark-up (interpretation); (2) the assignment of a well-defined meaning representation to concrete texts; and (3) issues of a community-wide consistent and accurate annotation of meaning.

By providing an essential, and heretofore non-existent, data set for training and evaluating natural language processing systems, the resultant annotated multilingual corpus of translations is expected to lead to significant research and development opportunities for machine translation and a host of other natural language processing technologies, including question answering (e.g., via paraphrase and entailment relations) and information extraction. Because of the unique annotation processes in which the each stage (IL0, IL1, IL2) provides a different level of linguistic and semantic information, a different type of natural language processing can take advantage of the information provided at the different stages. For example, IL1 may be useful for information extraction in question answering, whereas IL2 might be the level that is of most benefit to machine translation. These topics exemplify the research investigations that we can conduct in the future, based on the results of the annotation.

References

Baker, Collin and J. Fillmore and John B. Lowe, (1998). The Berkeley FrameNet Project. Proceedings of ACL.

Bateman, J.A., Kasper, R.T., Moore, J.D., and Whitney, R.A. (1989). A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model. Unpublished research report, USC/Information Sciences Institute, Marina del Rey, CA.

Carletta, J. C. (1996). Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2), 249-254

Dorr, Bonnie J. (2001). LCS Verb Database, Online Software Database of Lexical Conceptual Structures and Documentation, University of Maryland. http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html

Dorr, Bonnie J., (1993). *Machine Translation: A View from the Lexicon*, MIT Press, Cambridge, MA.

Farwell, David, and Steve Helmreich. (2003). Pragmatics-based Translation and MT Evaluation. In Proceedings of Towards Systematizing MT Evaluation. MT-Summit Workshop, New Orleans, LA.

Fellbaum, C. (ed.). (1998). WordNet: An On-line Lexical Database and Some of its Applications. MIT Press, Cambridge, MA.

Ferro, Lisa, Inderjeet Mani, Beth Sundheim and George Wilson. (2001). TIDES Temporal Annotation Guidelines. Version 1.0.2 MITRE Technical Report, MTR 01W0000041

Fillmore, Charles. (1968). The Case for Case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1--88. Holt, Rinehart, and Winston.

Fleischman, M., A. Echihabi, and E.H. Hovy. (2003). Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked. Proceedings of the ACL Conference. Sapporo, Japan.

Habash, Nizar and Bonnie Dorr. (2002). Interlingua Annotation Experiment Results. AMTA-2002 Interlingua Reliability Workshop. Tiburon, California, USA.

Habash, Nizar, Bonnie J. Dorr, and David Traum, (2002). "Efficient Language Independent Generation from Lexical Conceptual Structures," *Machine Translation*, 17:4.

Hajič, Jan; Vidová-Hladká, Barbora; Pajas, Petr. (2001): The Prague Dependency Treebank: Annotation Structure and Support. In Proceeding of the IRCS Workshop on Linguistic Databases, pp. .

Univesity of Pennsylvania, Philadelphia, USA, pp. 105-114.

Hovy, E.H., Philpot, A., Ambite, J.L., Arens, Y., Klavans, J., Bourne, W., and Saroz, D. (2001). Data Acquisition and Integration in the DGRC's Energy Data Collection Project, in Proceedings of the NSF's dg.o 2001. Los Angeles, CA.

Jackendoff, Ray. (1972). Grammatical Relations and Functional Structure. *Semantic Interpretation in Generative Grammar*. The MIT Press, Cambridge, MA.

Kingsbury, Paul and Martha Palmer and Mitch Marcus , (2002). Adding Semantic Annotation to the Penn TreeBank. Proceedings of the Human Language Technology Conference (HLT 2002).

Kipper, Karin and Martha Palmer (2000). Representation of Actions as an Interlingua. Proceedings of the Third AMTA SIG-IL Workshop on Interlinguas and Interlingual Approaches, Seattle, WA, April 30.

Knight, K., and I. Langkilde. (2000). Preserving Ambiguities in Generation via Automata Intersection. American Association for Artificial Intelligence conference (AAAI).

Knight, K, and Luk, S.K. (1994). Building a Large-Scale Knowledge Base for Machine Translation. Proceedings of AAAI. Seattle, WA.

Levin, Beth. (1993) "English Verb Classes and Alternations: A Preliminary Investigation", University of Chicago Press, Chicago, IL.

Levin, B. & Rappaport-Hovav, M. (1998). From Lexical Semantics to Argument Realization. Borer, H. (ed.) *Handbook of Morphosyntax and Argument Structure*. Dordrecht: Kluwer Academic Publishers.

Mahesh, K., and Nirenberg, S. (1995). A Situated Ontol-ogy for Practical NLP, in Proceedings on the Work-shop on Basic Ontological Issues in Knowledge Sharing at IJCAI-95. Montreal, Canada.

Mitamura, T., E. Nyberg, J. Carbonell. (1991). An Efficient Interlingua Translation System for Multilingual Document Production, in Proceedings of the Third Machine Translation Summit. Washington, DC.

Nyberg, E., T. Mitamura, K. Baker, D. Svoboda, B. Peterson, J. Williams. (2002) Deriving Semantic Knowledge from Descriptive Texts using an MT System. Proceeding of the 2002 Conference, Association for Machine Translation in the Americas.

Ogden, B., J. Cowie, E. Ludovik, H. Molina-Salgado, S. Nirenburg, N. Sharples and S. Sheremtyeva (1999). CRL's TREC-8 Systems: Cross-Lingual IR and Q&A, Proceedings of the Eighth Text Retrieval Conference (TREC-8).

Philpot, A., M. Fleischman, E.H. Hovy. (2003). Semi-Automatic Construction of a General Purpose Ontol-ogy. Proceedings of the International Lisp Confer-ence. New York, NY. Invited.

Stowell, T. (1981). Origins of Phrase Structure. *PhD thesis*, MIT, Cambridge, MA.

Tapanainen, P. and T Jarvinen. (1997). A non-projective dependency parser. In the 5th Conference on Applied Natural Language Processing / Association for Computational Linguistics, Washington, DC.

White, J., and T. O'Connell. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proceedings of the 1994 Conference, Association for Machine Translation in the Americas

Walker, Kevin, Moussa Bamba, David Miller, Xiaoyi Ma, Chris Cieri, and George Doddington (2003). Multiple-Translation Arabic Corpus, Part 1. Linguistic Data Consortium (LDC) catalog number LDC2003T18 and ISBN 1-58563-276-7

Verb Classification – Machine Learning Experiments in Classifying Verbs into Semantic Classes

Bart Decadt and Walter Daelemans

Center for Dutch Language and Speech
Language Technology Group
University of Antwerp – Campus Drie Eiken
2610 Wilrijk – Belgium
{bart.decadt,walter.daelemans}@ua.ac.be

Abstract

This paper presents the results of our machine learning experiments in verb classification. Using Beth Levin's semantic classification of the English verbs as a gold standard, we (i) test the hypothesis that the syntactic behavior of a verb can be used to predict its semantic class, and (ii) investigate whether a robust shallow parser can provide the necessary syntactic information. With 277 verbs belonging to six of Levin's classes, we do *type classification* experiments using RIPPER, an inductive rule learner. Having only a set of n most likely subjects or objects as features, this machine learning algorithm is able to predict the correct class with $\pm$ 58% accuracy. This result is comparable with results from other researchers, like Merlo and Stevenson, Stevenson and Joanis, and Schulte im Walde.

1. Introduction

1.1. Overview

In this paper, we present the results of our machine learning experiments in verb classification. We will start by sketching the background of this line of research, starting with Beth Levin's manual classification of the English verbs (Levin, 1993), and linking our work to it. Next, we will show the gold standard – six classes from Beth Levin's verb classification – used for evaluating the outcome of our experiments. We continue by explaining how we gathered data from the British National Corpus (BNC), and how we presented the data to the machine learning algorithm (RIPPER) used in the experiments. Finally, we report and analyze the results, compare them with related work and present the lines of research we will follow in the near future.

1.2. Background

In 1993, Beth Levin published her Ph.D. thesis (Levin, 1993), in which she described her handcrafted semantic classification of the English verbs. Her – very simplified – hypothesis is that the semantics of a verb determine to a large extent its syntactic behavior. By analyzing the English verbs along some syntactic criteria – among others the sub-categorization frames in which the verbs appear – she manages to distinguish 49 semantically coherent classes.

Levin's work was a source of inspiration, and a possibility for evaluation, for computational linguists working on semantic (verb) classification. The main goals of this line of research are (i) trying to classify or cluster words – in this case verbs – automatically according to their semantics, and (ii) determining which features are informative for this task.

Research on verb classification will enable us to do *lexical acquisition* for verbs: it will help in making or extending a lexicon with, for example, information on the semantic class of a verb. Another possible benefit of verb classification is that these techniques will help us to decide on the sub-categorization frame, or other syntactic or semantic information, of *unknown* or *new* verbs.

1.3. Our Research

In our research, we aim at *type classification* of English verbs into Levin's classes. With type classification, we mean that we collect information for the verbs, and for each particular verb we merge this information into one data vector. Then, on the basis of the collected information, we try to predict the semantic class of an unknown or new verb.

The information we use to classify the verbs is provided by a *shallow parser*: in the experiments reported here, we limited the information to the subjects and objects of the verbs. This information is fairly easy to extract from a shallow parsed corpus: we did not need to develop (complex) heuristics.

2. The Gold Standard

From Levin's classification, we selected a subset of 6 classes, some of which are divided in subclasses. These classes contain 318 verbs, of which we used only 277, because for the remaining 41 verbs we did not find or found not enough data in our corpus. Some of these verbs are ambiguous and appear in two of the six classes. For practical reasons, we ignored this ambiguity in our experiments: we assume that the 'main' class of a verb is the first class it appears in.

The selected classes and subclasses, with the number given by Levin to that class between brackets, are:

- *verbs of contact by impact* (18), containing four subclasses:
    - *hit verbs* (18.1)
    - *swat verbs* (18.2)
    - *spank verbs* (18.3)
    - *non-agentive verbs of contact by impact* (18.4)
- *poke verbs* (19)
- *verbs of contact* (20)
- *verbs of cutting* (21), containing two subclasses:

  - *cut verbs* (21.1)
  - *carve verbs* (21.2)
- *verbs of combining and attaching* (22), containing five subclasses:
  - *mix verbs* (22.1)
  - *amalgamate verbs* (22.2)
  - *shake verbs* (22.3)
  - *tape verbs* (22.4)
  - *cling verbs* (22.5)
- *verbs of separating and disassembling* (23), containing four subclasses:
  - *separate verbs* (23.1)
  - *split verbs* (23.2)
  - *disassemble verbs* (23.3)
  - *differ verbs* (23.4)

Table 1 shows the distribution of the 318 verbs over the 6 classes and 17 subclasses, expressed in numbers and percentages, and also lists some example verbs. Some classes are very small, like class 19, 22.5 and 23.4: machine learning algorithms can be expected to have difficulties learning these classes.

We evaluated our machine learning experiments in two ways. A first evaluation was done by looking at only the main classes: we will call this the *coarse-grained evaluation*. A second evaluation was done by taking the subclasses into account: we will call this the *fine-grained evaluation*.

From Table 1, we can induce a *random* and *default* baseline to compare our results with. The random baseline result is obtained by assigning class labels to the verbs according to the distribution in Table 1: in the coarse-grained case this results in 29.8% accuracy and in the fine-grained case in 9.4% accuracy. For the default baseline, we label each verb with the most frequent class label: this is class 22 in the coarse-grained case, resulting in 43.3% accuracy, and class 22.4 in the fine-grained case, resulting in 16.2% accuracy.

3. Data Acquisition and Representation

For the 277 verbs in the six classes from Levin, we collected information in the written part of the BNC (± 60M words). This corpus was shallow parsed with a memory-based shallow parser (Buchholz et al., 1999; Daelemans et al., 1999), developed at our research site[1]. After shallow parsing, we were able to make two lists for each verb: one with all the head nouns of the subjects and one with all the head nouns of the objects. These two lists were sorted by the statistical measure *likelihood ratio*: with this measure, the following two hypotheses for a subject-verb or object-verb pair are examined – see also (Manning and Schütze, 1999):

- Hypothesis 1 is the formulation of independence: the fact that the noun occurs in the subject position is not heavily determined by the verb.

$$H_1 : P(\text{noun as subject}|\text{verb}) = p = P(\text{noun as subject}|\neg\text{verb})$$

- Hypothesis 2 is the formulation of dependence: the fact that the noun occurs in the subject position is to a large extent determined by the verb.

$$H_2 : P(\text{noun as subject}|\text{verb}) = p_1 \neq p_2 = P(\text{noun as subject}|\neg\text{verb})$$

The values for p, p_1 and p_2 are computed as follows:

$$s = f(\text{noun as subject})$$

$$sv = f(\text{noun as subject,verb})$$

$$v = f(\text{verb}) \qquad V = f(\text{all verbs})$$

$$p = \frac{s}{V} \qquad p_1 = \frac{sv}{v} \qquad p_2 = \frac{s - sv}{V - v}$$

Assuming a binomial distribution:

$$b(k; n, x) = \binom{n}{k} x^k (1 - x)^{(n-k)} \tag{1}$$

the likelihoods of the two hypotheses above for the counts for s, v and sv attested in the BNC, are:

$$L(H_1) = b(sv; v, p)b(s - sv; V - v; p) \tag{2}$$
$$L(H_2) = b(sv; v, p_1)b(s - sv; V - v; p_2) \tag{3}$$

The log of the likelihood ratio can then be computed as follows:

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)} \tag{4}$$

$$\log \lambda = \log \frac{b(sv; v, p)b(s - sv; V - v; p)}{b(sv; v, p_1)b(s - sv; V - v; p_2)} \tag{5}$$

$$\log \lambda = \log L(sv; v, p) + \log L(s - sv; V - v; p)$$
$$- \log L(sv; v, p_1) \tag{6}$$
$$- \log L(s - sv; V - v; p_2) \tag{7}$$

where: $L(k, n, x)$ is equal to $x^k (1 - x)^{n-k}$.

The collected data was presented to the machine learning algorithm as follows: for each verb, we have only two features. The first feature is the *n most likely head nouns in the subject position* of the verb, and the second feature is the *n most likely head nouns in the object position of the verb*. The variable n ranged from 5 to 25, in steps of 5. With *n most likely* we actually mean *the at most n most likely* subjects or objects. If we only find 10 different head nouns in the subject or object position of some verb, we still include it in our experiments where the variable n is larger than 10.

We conclude this section with Table 2, in which we list some verbs with their 5 most likely (according to *likelihood ratio*) subjects and nouns, to illustrate how we presented our data to the machine learning algorithm RIPPER. Table 2 also shows that verbs from the same semantic class (can) have some nouns in common in their list of most likely subjects or objects.

[1] The shallow parser was developed in co-operation with the ILK research group from the University of Tilburg (The Netherlands).

class	# verbs	%	subclass	# verbs	%	examples
18	70	21.3%	18.1	24	7.3%	*beat knock*
			18.2	11	3.4%	*bite shoot*
			18.3	25	7.6%	*flog belt*
			18.4	10	3.0%	*crash thud*
19	6	1.8%	/	/	/	*poke stick*
20	12	3.7%	/	/	/	*kiss lick*
21	42	12.8%	21.1	10	3.0%	*hack slash*
			21.2	32	9.8%	*chop squash*
22	142	43.3%	22.1	15	4.6%	*blend link*
			22.2	42	12.8%	*unify pair*
			22.3	29	8.8%	*roll splice*
			22.4	53	16.2%	*string knot*
			22.5	3	0.9%	*cleave cling*
23	56	17.1%	23.1	12	3.7%	*divide part*
			23.2	13	4.0%	*break pull*
			23.3	29	8.8%	*unzip unlace*
			23.4	2	0.6%	*differ diverge*

Table1. The distribution of the verbs over the (sub)classes

verb	5 most likely subjects	5 most likely objects	main class label
pound	heart head foot rain blood	pavement stair earth road head	CLASS_18
drum	finger heart rain roar blood	finger business support interest heel	CLASS_18
chop	tbsp onion stir mushroom wash	parsley onion tomato garlic herb	CLASS_21
slice	blade onion oz carrot pain	bread tomato onion mushroom loaf	CLASS_21
seal	fate police lip door end	fate envelope victory gap deal	CLASS_22
clamp	hand finger car police mouth	hand teeth lip technique jaw	CLASS_22

Table2. Some examples of verbs with their 5 most likely subjects or objects.

4. Machine Learning Experiments

The machine learning algorithm we have experimented with is called RIPPER. RIPPER is an inductive rule learner: it induces classification rules from labeled examples by iteratively growing and then pruning rules. For more details on this algorithm, we refer to (Cohen, 1995) and (Cohen, 1996).

The advantage of using RIPPER is that it allows set-valued attributes: you do not need to convert the set-valued features to a binary format. Set-valued attributes is exactly what we are using: the feature *n most likely subjects* is the *set* of nouns appearing as head of the subject.

For each value of n, we searched the optimal parameter setting for this machine learning algorithm by doing leave-one-out training and testing: each one of the 277 verbs acted as test material, while the remaining 276 verbs were used as training material.

Depending on the type of features used – nominal, numeric, set-valued – RIPPER learns rules of the form "*if value for feature X (matches|contains|is greater than|is lesser than|...), then assign class label Y*". Below are two examples of rules – related to the verbs in Table 2 – learned by RIPPER from our dataset:

- `CLASS_21 4 0 IF OBJS ~ onion .`
- `CLASS_18 5 1 IF SUBJS ~ heart .`

We use nominal set-valued features, so these rules must be interpreted as "*if the set of n most likely objects contains onion, then assign class label CLASS_21*", and "*if the set of n most likely subjects contains heart, then assign class label CLASS_18*", respectively[2].

5. Results and Analysis

set-size	default setting	best setting	default baseline	random baseline
5	51.6	53.8	43.3	29.8
10	54.5	56.7		
15	53.4	54.2		
20	51.3	57.8		
25	52.7	56.7		

Table3. Coarse-grained evaluation results – accuracy in percentages

Table 3 shows the classification results of RIPPER, evaluated in the coarse-grained way. The numbers are accuracies expressed in percentages. The column *set-size* indicates the number of most likely subjects or objects we have used in the set-valued attributes for each verb. Though the

[2] The two pairs of numbers in these rules (`4 0` and `5 1`) indicate the number of data points in the training set to which the rule applies: the first number in the pair is the number of data points for which the rule predicts the class correctly, the second number is the number of data points to which the rule assigns an incorrect class label.

accuracies are not very high, in all cases the default setting scores better than both baseline results. With parameter optimization, we can improve the results a bit: the best result is obtained when the set-size is 20, yielding a classification accuracy of 58%.

Table 4 shows the results of RIPPER when analyzed in a fine-grained manner. It is clear that this task is much more difficult – but again all results with RIPPER's default settings are better than both baseline results. After parameter optimization and with a set-size of 15, the highest accuracy obtained is 31%.

set-size	default setting	best setting	default baseline	random baseline
5	23.1	25.6	16.2	9.4
10	26.7	28.5		
15	25.6	31.4		
20	24.6	31.1		
25	23.1	30.3		

Table4. Fine-grained evaluation results – accuracy in percentages.

In both evaluation types, the results are better than the baseline results, though the error reduction in the coarse-grained case is higher than in the fine-grained case. In the coarse-grained evaluation, the error reduction compared to the default baseline result is 23.6% and to the random baseline result is 39.8%. In the fine-grained case, the error reduction is 17.7% compared to the default and 24.3% compared to the random baseline.

	18	19	20	21	22	23
prec.	58.5	0.0	66.7	75.0	57.4	33.3
rec.	45.3	0.0	72.7	25.0	90.6	7.0
$F_{B=1}$	55.3	/	67.8	53.6	62.0	19.0

Table5. Precision, recall and $F_{B=1}$ scores for the six main classes.

Table 5 shows the precision, recall and $F_{B=1}$ scores for the six main classes in the best output we obtained with RIPPER in the coarse-grained evaluation. For most classes, precision is acceptable, but recall is quite low – exceptions are class 19 and 23. The reasonable precision but low recall suggests that for most classes, RIPPER learns a few rules which work well for a small set of verbs, but not for the whole class. The results for class 19 are very bad: it has zero precision and recall. Containing only 6 verbs, this class is the smallest: RIPPER does not have a lot of training material for this class. If we leave out during evaluation the classes with fewer than 10 verbs, which are class 19, 22.5 and 23.4, the classification accuracy improves a bit: 59% in the coarse-grained and 33% in the fine-grained case.

For class 22, recall is very high: more than 90%. This is because it is the *default class* for RIPPER: the machine learning algorithm starts by making rules for the smallest class first, then for the second smallest, and so on. For the largest class, there are no rules: if a new verb has to be classified, and all rules fail, RIPPER assigns it the label of the majority class.

The results in Tables 3, 4 and 5 indicate that to a certain extent, we can predict semantic classes from text with a machine learning algorithm by using little information provided by a shallow parser. For the coarse-grained case the results are reasonable, but for the fine-grained case we probably need more or other features.

6. Related Work

Table 6 summarizes very briefly the work of other researchers in the area of verb classification. The main difference between our research and the work summarized in Table 6 is that we have used nominal values, selected with a statistical criterion, whereas other researches have used numeric values – frequencies or probabilities.

The most work has been done by Merlo and Stevenson (see (Merlo and Stevenson, 2001; Stevenson and Merlo, 1999; Stevenson et al., 1999; Stevenson and Merlo, 2000; Merlo et al., 2002): with a decision tree learner and with frequency counts for five features, they obtain 69% classification accuracy. However, they classify verbs in only three classes which are not really semantically coherent and which do not correspond to classes from Beth Levin's classification.

In further research, Stevenson, in joint work with Joanis (Stevenson and Joanis, 2003), did use Levin's classes to evaluate the verb classification results: using a feature selection algorithm, which has to select among 220 features, and a decision tree learner, the best result they obtain is 58%. They also experimented with unsupervised learning, but results are much lower: their hierarchical clustering algorithm is able to reconstruct Levin's classification with 29% accuracy.

The state-of-the-art research comes from Schulte im Walde (Schulte im Walde, 1998): using frequency counts of verbs for a set of sub-categorization frames, she is able to reconstruct Levin's classification with unsupervised machine learning algorithms with 61% accuracy. She also did classification experiments with German verbs, using similar sub-categorization information (Schulte im Walde and Brew, 2002), but unfortunately she did not report the results in terms of classification accuracies.

Making a sound comparison of our results with the above mentioned research is not easy: they all use different classes and different machine learning methods. Moreover, it is never very clear whether the reported results are at the coarse-grained or at the fine-grained level. Still, we feel that our research can be best compared with Stevenson and Joanis' research – we even obtain similar results, 58% accuracy.

7. Future Work

In the following paragraphs we will briefly discuss our plans for near future work within the field of verb classification.

authors	classes	features	algorithm	result
Merlo and Stevenson	3 (Levin classes)	freq. counts for 5 features	C5.0	69%
Joanis and Stevenson	13 Levin classes	freq. counts for 220 features	C5.0	58%
			hierarchical clustering	29%
Schulte im Walde	30 Levin classes	freq. of verb with sub-categorization frames	iterative clustering	61%
			latent class analysis	54%

Table6. A summary of related work.

class	subclasses
9	9.1-6 (other verbs of putting) 9.7 (spray/load verbs) 9.8 (fill verbs)
10	10.1, 10.5 (steal and remove verbs) 10.4.1-2 (wipe verbs) 10.6 (cheat verbs)
13	13.1, 13.3 (recipient verbs)
26	26.1, 26.3 (benefactive verbs) 26.1, 26.3, 26.7 (object-drop verbs)
31	31.1 (amuse verbs) 32.2 (admire verbs)
43	43.2 (sound emission verbs)
45	45.1-4 (change of state verbs)
51	51.3.2 (run verbs)

Table7. Levin's classes used in Stevenson and Joanis' experiments.

Comparison with other work. First of all, to make a sound comparison with other researchers' results, we will do similar experiments using the verbs used in Stevenson and Joanis' experiments (Stevenson and Joanis, 2003). The classes to which these verbs belong are listed in Table 7. The class labels between brackets in this table are Stevenson and Joanis' interpretation of Levin's classes. The granularity of this classification is somewhere in between what we've called coarse- and fine-grained.

More features. We will also try to add more features which a shallow parser can provide, like for example the prepositions following a verb and the list of nouns in the prepositional phrase, and do similar experiments to find out whether these features can contribute to verb classification.

Token-based verb classification. Our verb classification experiments reported in this paper were *type-based*: information is collected by looking at individual tokens of a verb in a corpus, and for each verb, this information was collapsed in one data vector. It is interesting to investigate whether a *token-based* approach will also be successful at classifying verbs. The experimental set-up will then be as follows: for each token of a verb in a set of n verbs, a vector with information from a shallow parsed corpus (nominal values such as Part-of-Speech, chunk and relation tags of the focus word and surrounding words) will be constructed. For testing/evaluating this approach, we will do some kind of *leave-one-out cross-validation*: we will use all vectors for the tokens of n-1 verbs as training material, and classify all vectors for the tokens of the remaining verb (the *unknown* verb). In this architecture, the semantic class of the *unknown verb* is the label that is most often predicted.

This work is planned for the near future, and the results will be presented and discussed at the workshop.

Acknowledgments

This research was carried out within the context of the SemaDuct project, sponsored by the *Fonds voor Wetenschappelijk Onderzoek – Vlaanderen* (Fund for Scientific Research – Flanders).

8. References

Buchholz, Sabine, Jorn Veenstra, and Walter Daelemans, 1999. Cascaded grammatical relation assignment. In *Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing*. University of Maryland, College Park, MD, USA: The Association for Computational Linguistics.

Cohen, William W., 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*. Tahoe City, CA, USA.

Cohen, William W., 1996. Learning with set-valued features. In *Proceedings of the 13th National Conference on Artificial Intelligence*. Portland, Oregon, USA: The American Association for Artificial Intelligence.

Daelemans, Walter, Sabine Buchholz, and Jorn Veenstra, 1999. Memory-based shallow parsing. In *Proceedings of the Conference on Natural Language Learning*. Bergen, Norway: The Association for Computational Linguistics.

Levin, Beth, 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL, USA: The University of Chicago Press.

Manning, Christopher D. and Hinrich Schütze, 1999. *Foundations of Statistical Natural Language Processing*, chapter 8. Cambridge, MA, USA: The MIT Press, 2nd edition.

Merlo, Paola and Suzanne Stevenson, 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3):373–408.

Merlo, Paola, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria, 2002. A multilingual paradign for automatic verb classification. In (The Association for Computational Linguistics, 2002), pages 207–214.

Schulte im Walde, Sabine, 1998. Automatic semantic classification of verbs according to their alternation behaviour. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung*, 4(3):55–96. Lehrstuhl für Theoretische Computerlinguistik, Universität Stuttgart.

Schulte im Walde, Sabine and Chris Brew, 2002. Inducing german semantic verb classes from purely syntactic subcategorisation information. In (The Association for Computational Linguistics, 2002), pages 223–230.

Stevenson, Suzanne and Eric Joanis, 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of the Conference on Natural Language Learning*. Edmonton, Canada: The Association for Computational Linguistics.

Stevenson, Suzanne and Paola Merlo, 1999. Automatic verb classification using grammatical features. In *Proceedings of the 9th Conference of The European Chapter of The Association for Computational Linguistics*. Bergen, Norway: The Association for Computational Linguistics.

Stevenson, Suzanne and Paola Merlo, 2000. Automatic lexical acquisition based on statistical distributions. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarland University, Saarbrücken, Germany: International Committee on Computational Linguistics.

Stevenson, Suzanne, Paola Merlo, Natalia Kariaeva, and Kamin Whitehouse, 1999. Supervised learning of lexical semantic verb classes using frequency distributions. In *Proceedings of the SIGLEX-99 Workshop: Standardizing Lexical Resources*. University of Maryland, College Park, MD, USA: The Association for Computational Linguistics.

The Association for Computational Linguistics, 2002. *Proceedings of the 40th Annual Meeting of The Association for Computational Linguistics*. University of Pennsylvania, Philadelphia, PA, USA.

Unsupervised Semantic Disambiguation

Roberto Basili, Marco Cammisa

University of Rome Tor Vergata,
Department of Computer Science, Systems and Production,
00133 Roma (Italy),
{basili,cammisa}@info.uniroma2.it

Abstract

Supervised approaches to semantic disambiguation (ranging from classification of words into semantic fields to word sense disambiguation) make critical use of lexical resources. However the cost of hand annotation in such frameworks make them very expensive as the efficiency of training is often counterbalanced by huge effort in annotation of large scale data sets. In this paper we report the experience made during the John Hopkins 2003 Summer Workshop "*Semantic Analysis Over Sparse Data*" to implement and evaluate an unsupervised approach to semantic tagging. The method we propose uses the Princeton Wordnet ((Miller, 1995)) and its lexical inheritance network for corpus-driven parameter estimation of a simple Bayesian tagger. Results suggest that although the achieved performance is still below the ones obtained by relying entirely on annotated material (i.e. a fully supervised Maximum Entropy model), the gap is not considerable (about 10%).

1. Introduction

Semantic Disambiguation refers to a variety of specific processes that range from word sense disambiguation to named entity classification. Lexical semantic resources always play a central role in SD. They provide usually a static source of information where the target classes inventory is derived. In other cases they are paradigms of lexical representation and inspire the disambiguation model themselves. Recent work has been carried out to exploit Wordnet (Miller, 1995) as a source of information. In ((Abney and Light, 1999)) the Wordnet hyponimy relation among senses (i.e. synsets) is modeled in a probabilistic setting: the traversing of the hierarchy is seen as a Markov process and this enables a variety of statistical inferences about lexical preferences and disambiguation. On a similar light more recent works ((Ciaramita et al., 2003)) apply a different learning technique over the Wordnet hierarchy structure to complement sense descriptions with hyperonim information in order to increase the accuracy of word sense disambiguation. A common feature of these studies is the role of the lexical hierarchy as the main trigger of the decision function, that is the critical source evidence for disambiguation.

In this paper we propose an unsupervised model for the empirical parameter estimation (i.e. lexical and contextual probabilities) based on a similarity measure over Wordnet. The basic assumption is that similar syntactic behaviour of words is in fact due to similiarity on a semantic ground. Similarity is derived from noun clusters whose members show analogous syntactic behaviour, e.g. are direct objects of the same verb. For example, the fact that the noun $patient$, in a medical corpus, is the direct object of a verb like *treat*, (as well as other words like *women*, *people*, *children* or *male*) suggests a preference of its $person/human$ meaning while discarding other admissible but rare senses like '*grammatical role*'. We could refer this hypothesis as "one sense for syntactic collocation" in line with previous successful works in the area ((Yarowsky, 1995)).

We can increase in fact our local probability, $p(human|patient, treat_{DO})$, according to accompanying nouns (i.e. *women*, *people*, $\ldots$). Moreover, we can derive the overall confidence in such a sense, i.e. $p(human|patient)$ by accumulating all the observed contexts of the word $patient$. Thus, by generalizing nouns within their specific grammatical contexts over a reference hierarchy we could bootstrap the statistical tagger without the need of training material. In this paper we present a simple bayesian tagger (Section 2.) and the applied estimation technique exploiting semantic generalization over Wordnet (Section 3.). The semantic similarity measure introduced in (Basili et al., 2004) and its application to the tagger will be then defined and evaluated (Section 3.2.).

2. Corpus-driven Statistical tagging

In this section, we consider the problem of how to tag words in a text context by means of a simple probabilistic model. In analogy with statistical NLP methods, we want to use a corpus to estimate probabilities of semantic tags for a target word tw given the evidence brought by a specific *local* context of tw. In probabilistic terms, let r_1, ..., r_n express the underlying context of tw (e.g. syntactic relations in which tw enters). Let Γ (e.g. $Liquid$) represent the generic semantic class target of the tagging task. Then the appropriate class Ω for the target context satisfies the following model:

$$\Omega = \arg max_{\Gamma}\ p(\Gamma|tw, r_1, ..., r_n) \qquad (1)$$

Equation 1 depends on a generic context $tw, r_1, ..., r_k$ and on the set of syntagmatic evidences available for tw. In other words each r_i expresses a dependency in which tw (e.g. *patient*) is involved as head (e.g. of an adjectival modifier $(control, patient)_{Adj\_NP}$) or as a modifier (e.g. verbal argument $(treat, patient)_{V\_DO}$). As the different relations r_i active in a context can, with a reasonable accuracy, be considered independent each other, we can consider the conditioning event $tw, r_1, r_2, ..., r_k$ as the joint of k independent events $(tw, r_1), ..., (tw, r_k)$, respectively.

This greatly simplifies equation 1 that can be mapped into the following model:

$$\Omega = \arg max_{\Gamma} \prod_{i=1}^{k} p(\Gamma|tw, r_i). \quad (2)$$

The probabilities $p(\Gamma|tw, r_i)$ express syntagmatic collocations: in terms of a dependency grammar, r_i is isomorphic to a pair (d_i, w_i), where d_i is a dependency relation (e.g. direct object) and w_i is the involved head (or modifier) word (e.g. a verb like *treat*). Each (tw, r_i) thus represents a grammatical trigram. The accuracy of their estimates is clearly depending on the reliability of the corpus observation of these syntactic co-occurrences. We thus need: (1) an accurate estimation method that improves on the simple counts of (syntactic) trigrams that are highly affected by data sparseness problems (Section 3.) and (2) an effective smoothing technique, applied to Eq. 2, able to increase the robustness over sparse phenomena. The back-off method used in this paper is the focus of the next section.

2.1. Tackling Data Sparseness

In general, most of the $tw, r_1, ..., r_k$ contexts in Eq. 1 are characterized by very low counts. Moreover, as a single syntactic relation r_i is in general a bigram, parameters in Eq. 2 are grammatical trigrams. Data sparseness is such that maximum likelihood estimates may be very unreliable. As early introduced in (Katz, 1987) *backing-off* is an effective and largely adopted estimation technique. The idea behind this approach is to switch (i.e. back-off) from an estimation model (i.e. counts of triples) to a simpler one (i.e. counts for pairs) whenever the estimates become unreliable. The basic parameters (Γ, tw, r_i) in the Eq. 2 can thus be defined by back-off, as follows:

$$p_{BO}(\Gamma|tw, r) = \begin{cases} \hat{p}(\Gamma|tw, r) & if\ C(\Gamma, tw, r) > K \\ \alpha p_{BO}(\Gamma|tw) + \\ \beta p_{BO}(\Gamma|r) & otherwise \end{cases} \quad (3)$$

where $\hat{p}()$ is the estimation that is different from the back-off probability, $p_{BO}()$.

In order to back-off, two estimations of lower order than trigrams are needed.

Syntactic estimates $p_{BO}(\Gamma|r)$

They expresses the preference a syntactic collocate r gives to the semantic label Γ as follows:

$$p_{BO}(\Gamma|r) = \begin{cases} \hat{p}(\Gamma|r)\ if\ C(\Gamma, r) > K_{synt} \\ \delta p(\Gamma)\ \ otherwise \end{cases} \quad (4)$$

Equation 4 expresses the preference of a given syntatic dependence (captured by r), that is a selectional constraint related to the head (noun or verb) corresponding to r (e.g. the verb $treat$ for $tw = patient$). Whenever this dependence is not frequent enough (or even when it cannot be reliably estimated from the corpus), then the general probability of the semantic class Γ is used instead. K_{synt} defines the threshold over which a given relation r is expected to produce a reliable estimate.

Lexical estimates $p_{BO}(\Gamma|r)$

A lexical probability expresses the association between tw and Γ. This association is null when Γ does not characterize any definition/sense for tw. Otherwise, it can be estimated from the corpus as follows:

$$p_{BO}(\Gamma|tw) = \begin{cases} \hat{p}(\Gamma|tw)\ if\ tw \in Dictionary \text{ and} \\ \qquad C(\Gamma, tw) > K_{lex} \\ \gamma p(\Gamma) \quad otherwise \end{cases} \quad (5)$$

The estimate $p(\Gamma|tw)$ is considered reliable when tw is enough frequent in the corpus. Notice that estimating it would require to disambiguate, a priori, all the occurrences of tw in the corpus. However, a more realistic estimate can be also obtained. We can estimate $p(\Gamma|tw)$ from the different syntactic relations r in which tw enters in the corpus. We will also return on this point in the next section.

The back-off $p_{BO}(\Gamma|tw, r)$ depends on at least two different estimates, i.e.: $p_{BO}(\Gamma|tw)$ and $p_{BO}(\Gamma|r)$. Their suitable combination is very difficult to fix, a priori, as a lexical property. In fact, tw may be rare in the corpus so that the estimate $\hat{p}(\Gamma|tw)$ is not reliable. Also r can be rare or be characterized by poor selectional preferences so that $\hat{p}(\Gamma|r)$ provides no cue. In Eq. 3 the two contributions are weighted in a linear combination and a possible definition of their weights is:

$$\alpha = \frac{C(tw)}{C(tw)+C(r)}, \beta = \frac{C(r)}{C(tw)+C(r)}$$

Note that when both $\hat{p}(\Gamma|r)$ and $\hat{p}(\Gamma|tw)$ are in fact unreliable, the linear combination produces $p(\Gamma)$ as expected.

3. A semantic similarity measure for unsupervised tagging

The target problem of the suggested model is how to robustly estimate the probabilities $\hat{p}(\Gamma|tw)$ and $\hat{p}(\Gamma|r)$ and $\hat{p}(\Gamma|tw, r)$. The main idea behind our approach is that it is feasible and effective to estimate these probabilities from the corpus, without relying on training (i.e. manually annotated data). In order to achieve this result we will exploit a large scale lexical hierarchy, i.e. Wordnet (Miller, 1995).

The idea is that **if** large evidence about a syntactic phenomenon r (e.g. direct objects of a given verb like *to_drink*) can be collected from the corpus, **then** several useful implications can be drawned. First, semantic preference criteria can be induced from the set of words in the same syntactic dependency r. As an example, *water* and *beer* are not semantically equivalent but are similar at a certain extent, i.e. both are `liquids`. r helps to determine the level of generalization required to find equivalence. It is also true that different dependencies r provide semantic cues local to each relation r. They are thus independent each other, as they emphasize independent aspects of a word meaning.

Second, these preferences can be interpreted as overall probabilities conditioned only to r. The previous example suggests high values for $p(\texttt{ liquid } |water, to\_drink)$ and lower ones for $p(\texttt{ artifact } |water, to\_drink)$: although *artifact* is thus a possible sense for $water$ (as *facility*, *water system*) clustering $water$ with $beer$ (they are both occurring with relation r) produces a disambigation effect. We thus derive probabilities $\hat{p}(\Gamma|tw, r)$ from the set of nouns that similarly to tw appear in the same relation r. This converges toward the set of locally valid senses that are likely to generalize to other identical contexts. Moreover, this method is more robust with respect to data sparseness as all the contexts involving r in the corpus are used. It is to be noticed that $\hat{p}(\Gamma|tw, r)$ are, under this perspective, derived from $\hat{p}(\Gamma|r)$. The preferences are a property of all the words (like tw) related to r within a lexical hierarchy: they produce only the *common* generalizations that are in fact fewer than word senses.

tw that have low frequencies in the corpus can be studied over reliably observed contexts r that are highly selectional. Single contexts in isolation are prone to sparse phenomenon, but by collecting all nouns that (like tw) enter in a specific syntactic relation r (even once in the corpus), we increase the reliability of each estimate. Finally, a simple estimate of the corpus probability of $\hat{p}(\Gamma|tw)$ is thus:

$$\hat{p}(\Gamma|tw) = \sum_r \hat{p}(\Gamma|tw, r)p(tw|r) \tag{6}$$

The next section defines the estimation of the parameters of the proposed method, i.e. $\hat{p}(\Gamma|tw, r)$ and $\hat{p}(\Gamma|r)$.

3.1. A "*syntactic*" measure for word similarity.

Every syntactic collocation, r, corresponds to a set of nouns W: they are all the nouns that co-occur with r in the corpus, i.e. nouns whose grammatical position satisfies r at least once. As an example, fixing the verb *to_drink*, its frequent direct objects, as found in the British National Corpus ((Burnage and Dunlop, 1992)), include nouns like: *water, beer, alchool*. Any noun $tw \in W$, is ambiguous in general. However, a preference can be assigned to those senses l that make tw mostly similar to the other members of W (i.e. other liquids). The availability of a taxonomy T allows to compute this kind of similarity according to the topological properties of T.

We use here a notion of similarity called *Conceptual Density*, early introduced in (Agirre and Rigau, 1996). The conceptual density is a measure of the quality of a given generalization (i.e. node) t in the hierarchy with respect to a set of nodes, i.e. senses of input words W. It is the *information density* of the subtree rooted at t, with respect to the target set W. Conceptual density tells us how much similar are words in W with respect to one of their common generalizations: usually every non trivial generalization in W (i.e. common to at least two member nouns) suggests a subset of W with a specific semantic interpretation. When the set W is generated by a syntactic relation r, the conceptual density captures the compatibility of senses l of $tw \in W$ with the suitable semantic interpretation of r, i.e. selectional constraints of r. Mapping the conceptual density (cd) into a probability local to W, produces a distribution over senses l of $tw \in W$ and it gives rise to an efficient estimation technique.

DEF (*Conceptual Density*). Given a syntactic collocation r, that corresponds to the set W of nouns tw, a synset s in Wordnet that subsumes N of the lemmas in W, then the conceptual density, $cd^{(r)}(s)$, of s with respect to r is defined as follows:

$$cd^{(r)}(s) = \frac{\sum_{i=0}^{h} \mu^i}{area(s)} \tag{7}$$

where :

– $area(s)$ is the number of nodes in the s subhierarchy, i.e. a static property of Wordnet.
– μ is the average number of sons per node (i.e. branching factor) of the subhierarchy rooted by s. Notice that when nodes s lie on unbalanced branches of the hierarchy, the value for μ can approach 1 and a specific treatment is needed.
– h is the estimation of the depth of a(n ideal) tree that represents the N nouns. Its actual value is estimated by:

$$h = \begin{cases} \lfloor log_\mu N \rfloor & \textbf{iff } \mu \neq 1 \\ N & \textbf{otherwise} \end{cases} \tag{8}$$

Equation 7 applies to any common generalization of nouns tw in the target set W. The problem is the combinatorial explosion: every word $tw, tw' \in W$ is ambiguous and most pairs (tw, tw') may show common but useless generalizations (i.e. to be substituted by more specific subsumers). In order to reduce the number of generalizations produced in output from W (i.e. the set of nouns in a relation r) we applied a greedy technique based on the notion of *useful generalizations*. Given the target set W, a *useful* generalization in W is a Wordnet synset s such that s is an hyperonim of at least two words in W. Let S denote the set of all possible useful generalizations,

$$S = \{s | s \text{ is a useful generalization for } W\}.$$

Among the subsets $S' \subset S$, we look to the set O that is maximal with respect to the cumulated conceptual density. Formally, we look to $O \subset S$ such that:

$$\forall S' \subset S \quad \textstyle\sum_{s \in O} cd^{(r)}(s) \geq \sum_{s \in S'} cd^{(r)}(s)$$

The optimal set O can be found by the following *greedy* algorithm in Table 1.

Table1. Greedy Algorithm for the maximally dense generalizations

1. $O = \emptyset$.
2. $S = \{s | s \text{ is a useful generalization for } W\}$.
3. Rank elements in S according to decreasing values of $cd^{(r)}(s)$.
4. While W and S are not empty
   (a) Remove the highest ranked element s from S
   (b) Let C be set of nouns in W whose senses l have s as hyperonim
   (c) $W = W - C$
   (d) If $C \neq \emptyset$ then $O = O \cup \{s\}$
5. Return(O).

The outcome of the algorithm in Table 1 is the set O of *the maximally dense generalizations of at least two words* in W. If a word tw has no such generalization, it will not be represented in the resulting set O

3.2. Estimating sense probabilities from *cd* scores

The maximally dense generalizations O for target sets W (derived from syntactic relations r) couple senses l of target words tw with conceptual density values. Every $s \in O$ is considered a good generalizations of a sense l if s is along a generalization path starting from l. A maximum likelyhood approach can be here applied as evidence for each s is usually obtained by traversing the hierarchy from senses l related to several nouns.

DEF (*Local probability Estimation*). Given senses $l_1, ..., l_k$ of a $tw \in W$, and let O be the set of maximally dense generalizations of W in Wordnet, the conceptual density defined in Eq. (7) provides a probability estimate of $\hat{p}(l_i|tw, r)$ local to r as follows:

$$\hat{prob}(l_i|tw, r) = \frac{\sum_{s \text{ hyperonims of } l_i} cd^{(r)}(s)}{CD(tw, r)} \qquad \forall i \qquad (9)$$

where

$$CD(tw, r) = \sum_j \sum_{s \text{ hyperonims of } l_j} cd^{(r)}(s)$$

Notice that (as suggested in previous section) this definition of $\hat{p}(l_i|tw, r)$ can be used in turn to estimate $\hat{p}(l_i|r)$ by enumerating the different nouns $tw \in W$. Moreover, by enumerating all r valid for one word tw also $\hat{p}(l_i|tw)$ can be estimated.

The gap between fine grained lexicalized preferences and coarse grained class preferences is a drawback of the proposed method. Estimation proceeds from Wordnet synsets while the target tagsets may well lie within a different semantic system. Especially, when no specific relationship exist between the target classes Γ and the Wordnet synsets l_i, an explicit mapping is needed. However, in general we can expect that general systems of semantic labels (semantic tagsets) are clearly related to Wordnet. Sometimes a good mapping *a priori* can be made available especially when the tagset has been derived from Wordnet (for example the top level, i.e. topmost nodes). In other cases, e.g. named entity categories, a precise mapping can be easily built as a fixed mapping between nodes high in the Wordnet hierarchy and the target classes exist.

4. Experiments

The model proposed in previous sections has been experimented on the annotated portion of the British National Corpus, as obtained during the John Hopkins 2003 Summer Workshop "*Semantic Analysis Over Sparse Data*"[1]. The statistical word tagging model defined in Eq. 3 with estimates according to Eq. 10 has been applied to test material and compared with the results of a supervised Maximum Entropy model . The aim of the experiments was twofold:

– Evaluate the accuracy of the model with respect to test data of different complexity: a specific Wordnet-based baseline has been here introduced for measuring the effective impact of the estimation methods without any bias over the LDOCE to Wordnet mapping.
– Evaluate the accuracy of the unsupervised method contrastively with the supervised technique made available by the BNC data

The source corpus is made of human annotated sentences extracted from the British National Corpus (Burnage and Dunlop, 1992) which contained $198,970$ target noun phrases (i.e. test and training cases). About 94% inter-annotator agreement has been measured over a significant subset of the cases where choices of at least two annotators were available. A first portion of about $13,097$ instances was set aside: we will refer to this set as the blind corpus ($Blind$). The remainder of the human annotated corpus has been used for training with a fixed test set hereafter called $Held - Out$.

The corpus has been annotated with coarse grain semantic category inspired by the system of semantic codes of the Longman Dictionary of Contemporary English (Procter, 1978). Examples of classes are *Human*, *Abstraction*, *Animal* or *Collective Human*.

Unsupervised experiments used Wordnet as a basic resource for estimating source probabilities and a simple probabilistic model to annotate the test cases. Supervised approaches used Maximum Entropy methods primarily over annotated data extended with the results of parsed data (e.g. modifying adjectives and/or verbal heads), or with topical information (e.g. the topics of the source documents).

4.1. Mapping Wordnet to Longman semantic labels.

In all the experiments dealt with broad semantic classes. Following the work early presented in (Basili et al., 2003), in (Basili and Cammisa, 04) a technique based on *conceptual density* has been applied to map entries of the Longman Dictionary into Wordnet generalizations. For each couple tw, Γ we made available a probability distribution $p(l_i|\Gamma, tw)$ among the senses l_i of the word tw. After the analysis of all the words tw in the dictionaries the reverse probabilty was built so that

$$\hat{p}(\Gamma|tw, l_i) = \psi \frac{\hat{p}(l_i|\Gamma, tw)\hat{p}(l_i, tw)}{\hat{p}(\Gamma, tw)} \qquad (10)$$

where ψ is a normalizing factor. As a result a probability is obtained for every class. In the table 2 we report two examples where probability distributions over Wordnet synsets give rise to probability distributions for Longman-like classes.

The above method implements the mapping between the two dictionaries. Further details of the method and the description of the experiments can be found in (Basili and Cammisa, 04). As expected, the estimates $\hat{p}(\Gamma|tw, r)$ are made available for all known words in the dictionary from the source estimates $\hat{p}(l_i|tw, r)$. The large scale experiments aiming to evaluate the overall tagging model are then discussed in the next section.

[1] see URL at:
`http://www.clsp.jhu.edu/ws2003/groups/sparse/`

Table2. Mapping probability distibutions from Wordnet synsets (WN) to Longman-like classes (LD): Examples of two distributions for the word dog.

WN senses	$\hat{p}(l_i\|tw)$	*LD Classes*	$\hat{p}(\Gamma_i\|tw)$
1. dog, domestic dog	1	Male Animal	1.0
2. frump, dog	0		
3. dog, (informal)	0		
4. cad, bounder...	0		
5. pawl, detent, ...	0		
6. andiron	0		
1. dog, domestic dog	0	Female and Human	0.5
2. frump, dog	1	Male and Human	0.5
3. dog, (informal)	0		
4. cad, bounder...	0		
5. pawl, detent, ...	0		
6. andiron	0		

Table3. Performance of the Unsupervised Tagger.

Tagging Algorithm	Blind	Held-Out
Pick the 1^{st} sense	68,74%	72,40%
Unsupervised Tagger (*argmax*)	81,05%	75,45%
Unsupervised Tagger (coverage of 1^{st} 2 senses)	95,17%	91,28%

4.2. Evaluation of the Performance of the unsupervised tagger

The entire training process for the unsupervised model can be summarized as follows:

1. The training corpus has been parsed and head-modifier syntactic couples (and triples), whose head is a target noun (for semantic tagging), have been extracted. In this phase *word-preposition-noun* triples (e.g. *to_drink-during-dinner*, *water-with-gas*) or *verb-direct_object*, *subject-verb* couples, e.g. *drink-water*, or *drink-beer*, and *boy-drink*, are derived.
2. Classes of nouns are derived by fixing the grammatical heads and syntactic relations r, e.g. $beer$ and $water$.
3. Estimation of lexical probabilities (for senses l of tw) of the different nouns have been carried out by using Eq. 9. Local probabilities $\hat{p}(l|tw, r)$ and global probabiltiies $\hat{p}(l|tw)$, $\hat{p}(l|r)$ are also derived in this phase
4. Estimation of class probabilities $\hat{p}(l|tw, r)$ and global probabilities $\hat{p}(l|tw)$, $\hat{p}(l|r)$ is then derived (see Eq. 10).

After training, tagging a target noun tw within an incoming, grammatically analyzed, sentence:

$$tw \quad r_1 \quad r_2 \quad ... \quad r_k...$$

is carried out by Eq. 1 and 2.

Results of the supervised methods (based on ME trained with topical information plus adjectival modifiers) were around 85% for the $Held-Out$ data set of about $99,000$ cases and 93,4% on the $13,097$ cases ($Blind$). Table 3 reports the results of the unsupervised tagger over the blind corpus and over the Held-out.

The assumed baseline is the algorithm that tags the corpus according to the first Wordnet sense: the sense assumed by the Wordnet authors as the most common for n is mapped into class probabilites via Eq. 10. The third row tells us the number of correct decisions when both the first two solutions are accepted.

The major outcome is that unsupervised methods, not making use of annotated examples, are below the accuracy of supervised techniques but they are viable as converging towards high levels of performance. It is to be noticed that no actual large scale experiment in sense disambiguation or acquisition of selectional restrictions for verb arguments has been shown to outperform the "Pick the 1^{st} Wordnet sense" baseline, while the unsupervised tagger is well above this heuristic. Further exploration should study combinations of the Wordnet-based approach with the annotated material. Weakly supervision can be obtained by seeding the process with a small number of annotated cases and then adding external evidence to bootstrap to larger scales.

5. References

Abney, S. and M. Light, 1999. Hiding a semantic hierarchy in a markov model.

Agirre, Eneko and German Rigau, 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*.

Basili, R. and M. Cammisa, 04. Yet another mapping from ldoce to wn. In *Submitted fro publication*.

Basili, R., M. Vindigni, and F. M. Zanzotto, 2003. Integrating ontological and linguistic knowledge for conceptual information extraction. In *Proceedings of the IEEE/WIC WI-2003, Conference on Web Intelligence*. Halifax, CA.

Basili, Roberto, Marco Cammisa, and Fabio Massimo Zanzotto, 2004. A similarity measure for unsupervsed semantic disambiguation. in proceedings of language resources and evaluation conference. In *Proceedings of the LREC 204, Conference*. Lisbon, Portugal.

Burnage, G. and D. Dunlop, 1992. Encoding the British National Corpus. In *Proceedings of the 13th International Conference on English Language Research on Computerised Corpora*.

Ciaramita, Massimiliano, Thomas Hofmann, and Mark Johnson, 2003. Hierarchical semantic classification: Word sense disambiguation with world knowledge. in proceedings of 18th international joint conference on artificial intelligence.

Katz, S., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transaction on Acoustics, Speech, and Signal Processing*.

Miller, George A., 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Procter, P., 1978. *Longman Dictionary of Contemporary English*. Essex: Longman Group.

Yarowsky, David, 1995. Unsupervised word sense disambiguation rivaling supervised methods. in proceedings of the meeting of the association for computational linguistics.

An Unsupervised WordNet-based Algorithm for Relation Extraction

Mark Stevenson

Department of Computer Science
University of Sheffield
Sheffield
S1 4DP, UK
marks@dcs.shef.ac.uk

Abstract

Several machine learning techniques have been applied to the named entity (NE) recognition problem. However, there has been less progress on the problem of identifying relations between them; an important process in Information Extraction. This paper presents an unsupervised algorithm based on WordNet which discovers relations between entities in text, including those which have been identified as NEs. Comparative evaluation with a previously reported approach shows that the algorithm presented here is in some ways preferable and that benefits can be gained from combining the approaches using cotraining.

1. Introduction

Information Extraction (IE) systems often perform well in a particular domain but may be difficult to port to a new one. Many IE systems are based on knowledge-engineering approaches and prove difficult to adapt to a new domain. For example, the University of Massachusetts entered a system for the third MUC which required around 1,500 person-hours of expert labour to adapt it to that extraction task (Lehnert et al., 1992). IE itself can be though of as, at least, two sub-tasks: named entity recognition and relation extraction. The first of these is the process of identifying every item of a specific semantic type in the text. For example in the sixth MUC the semantic types included PERSON, ORGANIZATION and LOCATION. The second stage, relation extraction, involves the identification of appropriate relations between these entities and their combination into templates. The majority of IE systems carry out the stages of NE recognition and relation extraction as separate processes.

A lot of research has recently been carried out in named entity recognition and it can now be largely viewed as a solved problem; systems which achieve accurate results have been reported and implementations of named entity identifiers are available freely on the internet (e.g. http://www.gate.ac.uk). Unsupervised approaches to the NE recognition problem were presented by (Riloff and Jones, 1999) and (Collins and Singer, 1999). However, attempts to automate the relation extraction task have been less successful. Approaches include (Soderland, 1999), (Chieu and Ng, 2002), (Chieu et al., 2003) and (Zelenko et al., 2003). However, they all relied on supervised learning techniques and, consequently, depend upon the existence of annotated training data. To our knowledge the only approach which has made use of unsupervised learning techniques for relation extraction was presented by (Yangarber et al., 2000). This paper presents an alternative unsupervised algorithm for identifying relations between entities which are relevant to a particular IE task. This approach is compared with the one presented by (Yangarber et al., 2000) and found to complement it.

The approach presented here is based on the assumption that it is possible to learn patterns which are suitable for an IE system by presenting the system with small set of patterns, or "seeds", which are indicative of the patterns of interest. These patterns can then be compared with others in the corpus and the most similar added to the set of seeds. The similarity between patterns is determined using existing lexical similarity measures based on the WordNet lexicon (Fellbaum, 1998).

The learning algorithm is presented in Section 2., this includes details of how the corpus is pre-processed and background information on lexical similarity measures. Section 3. describes an alternative approach to the relation extraction task. An evaluation regime is described in Section 4. and the results of a comparative evaluation presented in Section 5..

2. System Details

2.1. Document Processing

A number of processing stages have to be applied to the documents before the learning process can take place. Firstly, named entities are marked. (Section 4.2. describes how this was carried out on the corpora used for the experiments described later in this paper.) The corpus is then parsed to identify Subject-Verb-Object (SVO) patterns in each sentence. Parsing was carried out using a version of MINIPAR (Lin, 1999) which was adapted to process the named entities marked in the text. The dependency trees produced by MINIPAR are then analysed to extract the SVO-pattern. Each tuple consists of either two or three elements. Sentences containing intransitive verbs yield tuples containing two elements, the second of which is the verb and the first its logical subject. For example, the sentence "The player scored on his debut" would yield the tuple player+score. The first two elements of tuples from sentences containing transitive verbs are the same while the third position represents the verb's object. Active and passive voice is taken into account in MINIPAR's output so the sentences "The professor taught the class" and "The class was taught by the professor" would yield the same triple;

professor+teach+class. The indirect object of ditransitive verbs is not extracted; these verbs are treated like transitive verbs for the purposes of this analysis.

2.2. Semantic Similarity

The aim of our learning mechanism is to learn patterns which are similar to those known to be relevant. To do this we make use of work which has been carried out on measuring semantic similarity between words. We experimented with several semantic similarity metrics[1] and found that the method proposed by (Lin, 1998) was most suitable for our application. This method relies on assigning numerical values to each node in the WordNet hierarchy representing the amount of information they contain (a technique developed by (Resnik, 1995)). This value was known as *Information Content*(IC) and was derived from corpus probabilities, so, $IC(s) = -\log(\Pr(s))$. For two senses, s_1 and s_2, the *lowest common subsumer*, $lcs(s_1, s_2)$, is defined as the senses with the highest information content which subsumes both senses in the WordNet hierarchy. Lin used these elements to calculate the semantic similarity of two senses according to this formula: $sim(s_1, s_2) = \frac{2 \times IC(lcs(s_1,s_2))}{IC(s_1)+IC(s_2)}$

It is simple to extend the notion of similarity between a pair of word senses to similarity between two words, w_1 and w_2, by choosing the pair of senses which maximise the similarity score. More formally, let $S(w_1)$ represent the set of senses for word w_1, where $S(w_1) = \{s_{11}, s_{12}, ...s_{1|S(w_1)|}\}$, and $S(w_2)$ the senses of w_2 ($S(w_2) = \{s_{21}, s_{22}, ...s_{2|S(w_2)|}\}$). The similarity of words w_1 and w_2 is determined according to equation 1.

$$word\_sim(w_1, w_2) = \underset{\substack{1 \leq i \leq |S(w_1)| \\ 1 \leq j \leq |S(w_2)|}}{MAX} sim(s_{1i}, s_{2j}) \qquad (1)$$

In Section 2.1. it was mentioned that named entities are marked in the text and so will appear in the SVO tuples. For example, the sentence “Jones left London” would yield the tuple NAMPERSON+leave+NAMLOCATION. The NAMPERSON and NAMLOCATION identifiers we used to denote the name classes do not appear in WordNet and so it will not be possible to compare their similarity with other words. To avoid this problem these tokens are manually mapped onto the most appropriate WordNet synset. This process is not particularly time-consuming since the number of named entity types with which a corpus is annotated is usually quite small. For example, in the experiments described later in this paper just seven named entity types were used to annotated the corpus.

We now extend the notion of word similarity to one of similarity between SVO patterns. The similarity of a pair of patterns can be computed from the similarity between the words in each corresponding pattern slot. So, imagine that p_1 and p_2 are patterns consisting of m and n elements respectively (where $1 \leq n, m \leq 3$) and that the mth element of pattern p_1 is denoted by p_{1m}. Then the similarity can be computed from equation 2 in which $MAX(m, n)$ is the greater of the values m and n. Normalising the sum of the word similarity scores by the longer of the two patterns takes into account patterns of differing length.

$$psim(p_1, p_2) = \frac{\sum_{i=1}^{n} word\_sim(p_{1i}, p_{2i})}{MAX(m, n)} \qquad (2)$$

2.3. A Semantic-Similarity-based Learning Algorithm

This idea of pattern similarity can be used to create an unsupervised approach to pattern generation. By taking a set of patterns which represent a particular extraction task we can compute the similarity of other patterns. Those which are found to be similar can be added to the set of accepted patterns and the process repeated. Our system starts with an initial set of seed patterns which are indicative of the extraction task. The rest of the patterns in the document set are then compared against the seeds to identify the most similar. Some of the similar patterns are then accepted and added to the seed set and the process repeated with the enlarged set of accepted patterns. The decision to accept a pattern can be either completely automatic or can be passed to a domain expert to include human judgement. Several schemes for deciding which of the scored patterns to accept were implemented and evaluated although a description would be too long for this paper. For the experiments described later we used a scheme where the four highest scoring patterns whose score is within 0.95 of the best pattern are accepted.

We shall now explain the process of deciding which patterns are similar to a given set of currently accepted patterns in more detail. Firstly, our algorithm disregards any patterns which occur just once in the corpus. The remainder of the patterns are assigned a similarity score based on equation 3. The score of a candidate pattern is restricted to the subset of accepted patterns which are “comparable” to it, denoted by C in this equation. This is useful since a good candidate pattern may be very similar to some of the accepted patterns but not others. For the purposes of this algorithm two patterns are said to be close if they have the same filler in at least one slot, for example john+phone+mary and simon+phone would qualify as close.

$$score(p) = \frac{\sum_{c \in C} psim(c, p) \times conf(c)}{log(|C|) + 1} \qquad (3)$$

Equation 3 includes the term $conf(c)$, a value in the range 0 to 1 representing the system’s confidence in pattern c. Such a confidence score is necessary since it is inevitable that some patterns accepted during the learning process will be less reliable than the seed patterns. These patterns may in turn contribute to the acceptance of other less suitable patterns and, if this process continues, the learning process may be misled into accepting many unsuitable patterns. The approach used here to avoid this problem is to introduce

[1] Unfortunately space constraints do not allow the description of these experiments here. The experiments described in this paper make use of a publicly available implementation of Lin’s similarity metric made available by (Patwardhan and Pedersen, 2003).

a score for pattern confidence which is taken into account during the scoring of candidate patterns.

We can be reasonably sure that seed patterns will be suitable for a domain and therefore these are given a confidence score of 1. Each newly accepted pattern is assigned a confidence score based on the confidence of patterns already accepted.

The confidence of the patterns accepted during iteration $i + 1$ is based on the confidence of the patterns which contributed towards it acceptance (that is those which are in the set C in equation 3) and the confidence scores they had in the previous iteration. The formula for calculating the score is shown in equation 4.

$$conf^{i+1}(p) = \frac{\sum\limits_{c \, \epsilon \, C} conf^{i}(c)}{|C|} \cdot \left(\underset{c \, \epsilon \, C}{MAX} \sqrt{\frac{psim(p,c)}{psim(p,p)}} \right) \tag{4}$$

Equation 4 guarantees that the confidence of the newly accepted pattern will be no greater than the highest confidence score of the patterns which contributed to its acceptance. However, the confidence score of patterns which have already been accepted can also be improved if they contribute to a new pattern whose score is higher than their own. So, if $conf^{i+1}(p) > conf^{i}(c)$ for some $c \, \epsilon \, C$ in equation 4 then $conf^{i+1}(c)$ is increased to $conf^{i+1}(p)$.

3. Alternative Approaches

3.1. Distributional Similarity

The approach described in Section 2.3. was inspired by an unsupervised algorithm for learning relations described by (Yangarber et al., 2000) who presented an approach which can be thought of as *document centric*. It is motivated by the assumption that a document containing a large number of patterns which have already been identified as relevant is likely to contain further patterns which are relevant to the domain. This is in contrast to the approach presented here which assumes that new relevant patterns can be found by choosing ones which are semantically similar to those already identified and makes no reference to the notion of document relevance.

It is important to mention that the unsupervised learning algorithm based on distributional similarity used for these experiments is not identical to the one described by (Yangarber et al., 2000). That system makes some generalisations across pattern elements by grouping certain elements together. However, there is no difference between the expressiveness of the patterns learned by either approach.

3.2. Cotraining

Cotraining (Blum and Mitchell, 1998) is a technique which allows learning algorithms to work together by sharing their results. It operates by combining the sets of patterns returned after each iteration thereby allowing the generalisation algorithms to share information. A theoretical assumption behind co-training (Blum and Mitchell, 1998) is that the generalisation procedures are independent, which is the case here since they are based on different information sources: the distribution of patterns across documents in a corpus and the semantic information contained in WordNet.

In the original cotraining method presented by (Blum and Mitchell, 1998) two unsupervised algorithms collaborate by sharing their results. In the context of this paper this would mean that after each iteration the set of accepted patterns would be expanded to include all patterns learned by both the semantic- and distributional-based classifiers. We implemented this approach but found that it did not perform well. Instead we adopted a different from of cotraining which aimed to maximise the use of two classifiers by accepting the set of patterns which are proposed by both classifiers. In other words we accept the intersection of the two sets of patterns, rather than the union. If the intersection is empty each classifier adds the best pattern identified during the previous iteration.

4. Evaluation

(Yangarber et al., 2000) noted that quantative evaluation of pattern induction systems is difficult to achieve. The discovery process does not easily fit into MUC-style evaluations since the learned patterns do not directly fit into an IE system. However, in addition to learning a set of patterns, the system also notes the relevance of documents relative to a particular set of seed patterns. (Yangarber et al., 2000) quantatively evaluated the documents relevance scores. This evaluation is similar to the “text-filtering” sub-task used in MUC-6 in which systems were evaluated according to their ability to identify the relevant documents for the extraction task. A similar evaluation was implemented for this study which allows comparison between the results reported by (Yangarber et al., 2000) and those reported here.

4.1. Document Filtering

The evaluation measure used by (Yangarber et al., 2000) relies on each document in the collection being assigned a relevance score which represents the appropriateness of that particular document to the extraction task. This measure is also used to help decide which of the potential candidate patterns to accept. The learning system described in Section 2.3. does not require the notion of document relevance but it is possible to make use of the scheme used by (Yangarber et al., 2000) to generate document relevance scores based on the patterns which were accepted by our algorithm.

During the first iteration of the algorithm each document matched by a seed pattern is assigned a relevance score of 1 and all other documents are given a score of 0. On subsequent iterations each pattern is assigned a *precision score* based on the mean relevance of the documents which it matches, as shown in Formula 5.

$$Prec^{i+1}(p) = \frac{\sum\limits_{d \epsilon H(p)} Rel^{d}(d)}{|H(p)|} \tag{5}$$

The pattern precision scores are then used to update the document *relevance scores* using formula 6 where K is a subset of the set of accepted patterns which apply to the document d.

$$Rel^{i+1}(d) = \frac{\sum_{d \epsilon H(K)} Rel^i(d)}{|H(K)|} \quad (6)$$

(Yangarber et al., 2000) use the pattern precision score in their algorithm to decide which pattern to accept during the next iteration. However, all that is needed to make use of this mechanism is for the documents to have a relevance score and for the set of patterns to be extended at each iteration. This is true of our approach since the documents which are matched by seed patterns can be assigned a relevance score of 1 and the set of accepted patterns is provided by the WordNet-based algorithm described in Section 2.3..

After each iteration the document relevance scores can be used to determine how accurately the set of induced patterns can discriminate between documents which are relevant to the extraction task and those which are not. This is carried out by calculating precision and recall scores which take into account the document relevance score provided by the system. For example, if a relevant document is assigned a score of X then the count of correct classification is incremented by X and the incorrect classification count by $1 - X$.

4.2. Evaluation Corpus

The corpus used for the experiments was compiled from two sources: the training and testing corpus used in the sixth Message Understanding conference (MUC-6) (muc, 1995) and a subset of the Reuters Corpus (Rose et al., 2002). The MUC-6 task was to extract information about the movements of executives from newswire texts. A document is relevant if it has a filled template associated with it. 590 documents from a version of the MUC-6 evaluation corpus described by (Soderland, 1999) were used.

The documents which make up the Reuters corpus are also newswire texts. However, unlike the MUC-6 corpus they have not been marked for relevance to the MUC-6 extraction task. Each document in the Reuters corpus is marked with a set of codes which indicate the general topic of the story. One of the topic codes (`C411`) refers to management succession events and this can be used to identify relevant documents. A corpus of 6,000 documents was extracted from the Reuters corpus. One half of this corpus contained the first (chronologically) 3000 documents marked with the `C411` topic code and the remainder contained the first 3000 documents which were not marked with that code.

Each document in this corpus was preprocessed as outlined in Section 2.1.. Relevant named entities are already marked in the MUC corpus and, since these have been manually verified, they were used for the preprocessing. These simply had to be transformed into a format suitable for the adapted version of MINIPAR. Named entities are not marked in the Reuters corpus and so the 6,000 documents were run through the named entity identifier in GATE (Cunningham et al., 2002) before parsing.

The combined corpus consisted of 6,590 documents. This yielded 142,563 pattern tokens from a set of 89,342 types. 72,997 patterns appeared just once and these were effectively discarded since our learning algorithm only considers patterns which occur at least twice (see Section 2.3.).

5. Results

We experimented with the smx_sim unsupervised algorithm (described in Section 2.), the distributional similarity algorithm (Section 3.1.) and their combination using co-training. The set of seed patterns listed in Table 1. The seed patterns matched 556 documents with a precision and recall of 1 and 0.26 respectively.

NAMCOMPANY+appoint+NAMPERSON
NAMCOMPANY+elect+NAMPERSON
NAMCOMPANY+promote+NAMPERSON
NAMCOMPANY+name+NAMPERSON
NAMPERSON+resign
NAMPERSON+depart
NAMPERSON+quit
NAMPERSON+step-down

Table1. Seed patterns for the management succession domain extraction task

These approaches were compared against a baseline system, called "random", which randomly accepted four candidate patterns at each iteration.

The results of this experiment are shown in Table 2 which shows the precision, recall and F-measure scores for each approach. The leftmost column indicates the number of iterations for which each algorithm has run. Continuous F-measure scores are presented in graphical format in Figure 1.

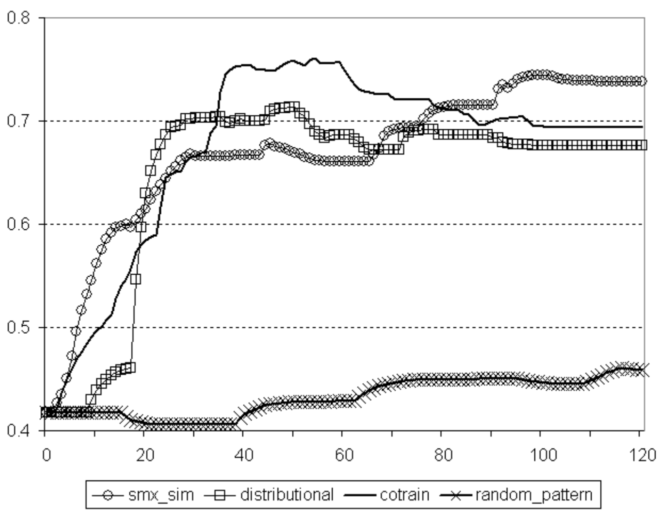

Figure1. F-measure scores for alternative approaches applied to the document filtering task over 120 iterations

It can be seen that each of the three methods outperforms the random baseline. The baseline method records a slight improvement in F-measure score during the learning

	random			smx_sim			distributional			cotraining		
#	P	R	F	P	R	F	P	R	F	P	R	F
0	1.00	0.26	0.42	1.00	0.26	0.42	1.00	0.26	0.42	1.00	0.26	0.42
20	0.89	0.26	0.41	0.73	0.53	0.61	0.73	0.55	0.63	0.83	0.45	0.58
40	0.88	0.27	0.42	0.62	0.72	0.67	0.67	0.74	0.70	0.76	0.75	0.75
60	0.88	0.28	0.43	0.60	0.74	0.66	0.59	0.83	0.69	0.69	0.81	0.75
80	0.89	0.30	0.45	0.63	0.83	0.71	0.57	0.87	0.69	0.58	0.93	0.71
100	0.85	0.30	0.45	0.63	0.91	0.74	0.55	0.87	0.68	0.55	0.94	0.69
120	0.82	0.32	0.46	0.62	0.91	0.74	0.55	0.87	0.68	0.55	0.94	0.69

Table2. Comparison of different approaches to document filtering task over 120 iterations

process. This is because the set of seed patterns matches few documents in the corpus which is split roughly 50/50 in terms of relevant and irrelevant documents. Therefore, there are more patterns which improve performance than hinder it.

The two learning algorithms, smx_sim and distributional, behave quite differently. The improvement of the smx_sim algorithm is slower than the distributional algorithm although the performance after 120 iterations is higher. The approach which records the highest score is the combination of these approaches using cotraining. Table 3 shows the best score recorded for each algorithm and the iteration during which it was recorded. The best score is recorded using cotraining after 53 iterations. The smx_sim algorithm achieves nearly as good an F-measure but takes twice the number of iterations required by cotraining. The cotraining and distributional algorithms record their highest F-measures with roughly equal precision and recall while the smx_sim algorithm record a very high recall at the expense of precision.

approach	P	R	F	#
smx_sim	62.3	90.5	73.8	106
distributional	66.1	77.2	71.3	49
cotraining	73.7	78.2	75.9	53

Table3. Best score recorded for each approach and iteration after which it was recorded

Table 4 shows the patterns learned by the cotraining approach after the first and tenth iteration. It can be seen that the quality of the patterns is quite mixed. Some of the patterns (e.g. `NAMCOMPANY+hire+NAMPERSON`) appear very relevant to the extraction task. While there are others (e.g. `NAMPERSON+begin`) which do not seem relevant. However, this may in part be due to the restricted representation of sentences used in this system; the relevance of these patterns may be more obvious given a richer representation. For example, the sentence “Jones begins his new role next month.” would produce the pattern `NAMPERSON+begin` and is relevant to the management succession task.

Patterns learned after first iteration:
NAMCOMPANY+hire+NAMPERSON
president+resign
cfo+resign
ceo+resign
NAMCOMPANY+say+NAMPERSON
NAMCOMPANY+be+NAMPERSON
Patterns learned after tenth iteration:
NAMPERSON+die
NAMPERSON+win
NAMPERSON+assume
NAMPERSON+become+NAMPOST
NAMPERSON+begin

Table4. Patterns learned by the cotrain approach after the first and tenth iterations

6. Conclusions

The approach presented here is inspired by the approach of (Yangarber et al., 2000) but makes use of a different assumption regarding which patterns are likely to be relevant to a particular extraction task. Evaluation showed that the proposed approach performs well when compared with the existing algorithm. In addition, the approaches are complementary and the best results are obtained when the results of the learning algorithms are combined using cotraining.

We plan to extend the work presented here in a number of ways including the exploration of richer sentence representation schemes such as the “Subtree model” (Sudo et al., 2003) and methods for deciding when to stop the learning process such as the counter-training approach proposed by (Yangarber, 2003).

Acknowledgements

The author is grateful to Roman Yangarber for advice on the implementation of the distributional similarity algorithm and to Neil Ireson and Angus Roberts from Sheffield University for comments on early drafts of this paper.

7. References

1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. MUC, San Mateo, CA: Morgan Kaufmann.

Blum, A. and T. Mitchell, 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the*

11th Annual Conference on Computational Learning Theory (COLT-98). Maddison, WI.

Chieu, H. and H. Ng, 2002. A Maximum Entroy Approach to Information Extraction from Semi-structured and Free Text. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence (AAAI-02)*. Edmonton, Canada.

Chieu, H., H. Ng, and Y. Lee, 2003. Closing the Gap: Learning-based Information Extraction Rivaling Knowledge-engineering Methods. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*. Sapporo, Japan.

Collins, M. and Y. Singer, 1999. Unsupervised models for Named Entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. College Park, MA.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan, 2002. GATE: an Architecture for Development of Robust HLT. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, PA.

Fellbaum, C. (ed.), 1998. *WordNet: An Electronic Lexical Database and some of its Applications*. Cambridge, MA: MIT Press.

Lehnert, W., C. Cardie, D. Fisher, J. McCarthy, E. Riloff, and S. Soderland, 1992. University of Massachusetts: Description of the CIRCUS System used for MUC-4. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*. San Francisco, CA.

Lin, D., 1998. An Information-Theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine learning (ICML-98)*. Madison, Wisconsin.

Lin, D., 1999. MINIPAR: a minimalist parser. In *Maryland Linguistics Colloquium*. University of Maryland, College Park.

Patwardhan, S. and T. Pedersen, 2003. The WordNet::Similarity package v0.06. http://www.d.umn.edu/˜tpederse/similarity.html.

Resnik, P., 1995. Using Information Content to evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*. Montreal, Canada.

Riloff, E. and R. Jones, 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*. Orlando, FL.

Rose, T., M. Stevenson, and M. Whitehead, 2002. The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02)*. La Palmas de Gran Canaria.

Soderland, S., 1999. Learning Information Extraction Rules for Semi-structured and free text. *Machine Learning*, 31(1-3):233–272.

Sudo, K., S. Sekine, and R. Grishman, 2003. An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*.

Yangarber, R., 2003. Counter-training in the discovery of semantic patterns. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*. Sapporo, Japan.

Yangarber, R., R. Grishman, P. Tapanainen, and S. Huttunen, 2000. Unsupervised discovery of scenario-level patterns for information extraction. In *Proceedings of the Applied Natural Language Processing Conference (ANLP 2000)*. Seattle, WA.

Zelenko, D., C. Aone, and A. Richardella, 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.

Exploiting the Semantic Fingerprint for Tagging ”Unseen” Words

Fabio Massimo Zanzotto and Armando Stellato

Department of Computer Science
University of *Roma*, *Tor Vergata*, Roma (Italy)
{zanzotto,stellato}@info.uniroma2.it

Abstract

In this paper we want to investigate the use of external and ”orthogonal” semantic resources in building coarse-grained semantic taggers. Our aim is to reduce the degree of supervision for the learning phase by keeping small the set of words whose behaviour has to be manually studied throughout a corpus. We introduce the notion of *semantic fingerprint* in order to exploit these external semantic resources in both machine learning and statistical models. Semantic fingerprints allow a straightforward integration of hierarchical information in the feature vector model. We will study and experimentally compare the effect on coarse-grained semantic taggers of different kinds of semantic fingerprints based on different semantic resources.

1. Introduction

Words seem to be *semantically conservative* as they tend to keep their preferred sense when taken in topically coherent document collections. This intuition underlies many studies in word sense disambiguation as (Madhu and Lytle, 1965; Gale et al., 1992; Resnik, 1997). Let us take, for instance, the famous example of the word *bank*. Even if this word is highly ambiguous, i.e. it has the senses of *institution*, *building*, and *river bank*, a semantic tagger can easily choose the correct sense if the knowledge domain is given. When dealing with texts related to financial news the most probable tag would be *institution*. On the other hand, whenever analysing navy related bulletins it is likely that the word assumes the *river bank* sense. There is some evidence for this phenomenon and it seems to be even very intense when coarse-grained semantic dictionaries are used. In the portion of the British National Corpus tagged with respect to a subset of the LDOCE categories the semantic tagging activity has a perplexity close to 1 (Guthrie et al., 2004).

Exceptions to the word attitude of being semantically conservative seem to be rare. Given the above, the best (and simplest) starting point in building a semantic tagger for a given knowledge domain seems to be collecting a good estimation of the prior distribution of word semantic tags in that specific domain. This estimation would require that, in a new domain, each word is observed and tagged in a sufficient number of instances in order to derive the most likely sense.

In this paper we investigate the possibility of reducing the words over which this manual tagging activity should be done. The manual semantic tagging done for a portion of the dictionary words in the domain corpus should be used to give hints to an automatic classifier in order to discover the most probable semantic tag for the remaining words. For instance, the preferred *investor* sense for the word *bear* in a financial domain (discovered and imposed by manually tagging word instances in the text collection) should help to deduce the same preference for the word *bull*. We claim that, when building a semantic tagger based on a coarse-grained semantic dictionary D, such a kind of beneficial effect may be obtained using a external and more fine-grained lexical resource D'. To investigate this claim we introduce the notion of *semantic fingerprint* as a way to exploit hierarchical semantic information in the classical machine learning feature vectors. After a short discussion on the envisaged procedure for building a semantic tagger (Sec. 2.), we will describe how the semantic fingerprint notion is useful for introducing hierarchical semantic knowledge in the classical feature vector model underlying many machine learning algorithms (Sec. 3.). Then, we will introduce the probabilistic classifiers used to investigate the usability of the semantic fingerprint when building semantic taggers (Sec. 4.). Finally, results of the experimental investigation are discussed (Sec. 5.).

2. Building a semantic tagger for a knowledge domain

The knowledge domain where words are used seems to give relevant hints to infer their sense. In early Machine Translation projects, this information was used to prepare *ad hoc* domain dictionaries containing only word senses relevant for the particular domain (e.g. (Oswald and Lawson, 1953)). Eliciting senses from the dictionary to build a domain sense tagger is not a perfect solution, as domain does not eliminate ambiguity for some words (as noticed in (Dahlgren, 1988)) and as some rare word senses may appear. However, it would be unreasonable not to take into consideration the bias induced by specific domains. For this reason, though all of the word senses have to be kept in our dictionary, domain sense preference for words should be included in a semantic tagger and used to modify sense distribution accordingly. Domain bias may be included in a probabilistic form.

In a closed world assumption, largely done in word sense disambiguation and in semantic tagging (Ide and Veronis, 1998), a dictionary D is used to describe all the necessary word senses. The prior distribution of senses for a word is generally uniform. The exploitation of the domain priming information requires therefore the re-estimation of the sense distribution for each word in the dictionary over the particular domain. As a knowledge domain is often represented as a coherent document collection, the sense distribution has to be estimated observing words in their context. This manual work should be done for each word in the dictionary that is likely to appear in the corpus. This activity will constitute the supervision for the semantic tagging building procedure.

In line with other approaches, our aim is to investigate procedures for building semantic taggers that are open to the reduction of the amount of supervision. Let us examine the general procedure for building a tagger in the closed world assumption. Given a semantic dictionary D with its semantic tag catalogue T and an unnannotated domain corpus C, the target of the procedure is to build the classification function $Tagger(i)$ that assigns the correct semantic tag $t \in T$ to each word instance i for the domain. The building model is the following:

- divide the dictionary D in two halves, namely $Train$ and $ToTag$
- annotate the occurrences of the words belonging to $Train$ in the corpus C
- train a classifier $Tagger$ on the instances of $Train$ in the corpus C
- tag the *unseen* word instances with the trained classifier $Tagger$, i.e. the instances of the $ToTag$ words in the the corpus C

It is worth noticing that the procedure has the ultimate aim to decide the preferred sense for each word (with respect to the semantic catalogue T) in the corpus C, that is representative of the target knowledge domain.

According to the desired degree of unsupervision, the first step of the procedure may be pursued in many ways. As a first possible choice, the dictionary may be randomly divided into two halves. In an active learning environment, the $Train$ section should include the most informative words, e.g. the most frequent words in the corpus C. Finally, in a completely unsupervised approach words in $Train$ may be the unambiguous words in the dictionary D while words in $ToTag$ are all the ambiguous ones. Ambiguity should be defined with respect to the target semantic dictionary. In this latter case the activity of tagging the word instances in the corpus C is eliminated.

In this general procedure, the real problem is to decide what information the classification algorithm $Tagger$ has to rely on. We will try to demonstrate that the use of lexical semantic resource D' other than the dictionary D helps in increasing the performances of the semantic tagger. In this paper we will focus only on information related to the word to be tagged, neglecting all the contextual evidence that could help in the disambiguation process.

Given therefore the external resource D' with its semantic tags T', our basic idea is that words appearing with a given frequency in the corpus shape the behaviour of the other words as some nodes in T' will be more active than the others. If T' is more fine-grained than S or represents an ”orthogonal” semantic model, it should help in classifying words with respect to T (see the above example between *bear* and *bull* in an financial domain).

3. Classification Function and Semantic Fingerprints

What we are seeking is a classification function $Tagger(i) = t$ that proposes a class t for any given instance i representing a word in a text. This classification function will observe objects in an instance space I assigning a class t in a set of possible categories T, i.e.:

$$Tagger : I \to T$$

In machine learning, this function assumes a variety of shapes, (e.g. decision trees in (Quinlan, 1993)), whereas in a probabilistic framework (e.g. the Maximum Entropy model (Jelinek, 1998)), it is seen as a selector of the most probable category given the conditions imposed by i, i.e.:

$$Tagger(i) = argmax_{t \in T} P(t|i)$$

Obviously, the categorisation is possible if some regularities appear in the space of the instances I. These regularities can be detected whenever observable features are defined. Given the observable features F_1,...,F_n, an instance $i \in I$ identifies a point in the space $F_1 \times ... \times F_n$, i.e.:

$$i = (f_1, ..., f_n) \in F_1 \times ... \times F_n$$

In machine learning this model is generally called feature-value vector and underlies many algorithms, as the ones gathered in (Witten and Frank, 1999).

With this general model in mind, we will try to describe in the rest of the section how an external semantic resource based on an hierarchical organisation can be used. We will firstly concentrate on the general limitations of the feature vector with respect to this problem and we will then propose a possible solution that we call *semantic fingerprint*.

3.1. Features of the feature vector

Many machine learning algorithms (as the ones in (Witten and Frank, 1999)) use the feature-value model assuming:

- the *a-priori independence*: each feature is *a priori* independent from the others and, therefore, no possibility is foreseen to make explicit relations among the features;
- the *flatness* of the set of the values for the features: no hierarchy among the values of the set is taken into consideration;
- the *certainty of the observations*: given an instance I in the feature-value space, only one value is admitted for each feature.

Under these limitations ML algorithms offer the possibility of selecting the most relevant features that may help in deciding whether or not an incoming object in the feature-value space is instance of a given concept.

Exploiting the feature-value vector model and the related learning algorithms in the context of natural language processing may then be a very cumbersome problem, especially when the successful bag-of-word abstraction (Salton and Buckley, 1988) is abandoned for deeper language interpretation models. The a-priori independence among features, the flatness of the values, and the certainty of the observations are not very well suited for syntactical and semantic models. On the one side, syntactical models would require the possibility of defining relations among features in order to represent either constituents or dependencies among words. On the other side, a semantic interpretation of the words (intended as their mapping in an is-a hierarchy such as WordNet (Miller, 1995)) would require the possibility of managing hierarchical value sets in which the substitution of a more specific node with a more general one can be undertaken as a generalisation step. Finally, the ambiguity of the interpretations (either genuine or induced by the interpretation model) stresses the basic assumption of the *certainty of the observations*. Due to ambiguity, a given instance of a concept may be seen in the syntactic or the semantic space as a set of alternative observations. The limits of the underlying interpretation models in selecting the best interpretation requires specific solutions to model *uncertainty* when trying to use feature-value-based machine learning algorithms for learning concepts represented by natural language expressions.

3.2. Hierarchies in the Feature Vector: the Semantic Fingerprint

The use of a hierarchical lexical resource is really cumbersome especially when coupled with the uncertainty of the observations. If we want to rely on an external semantic resource, we surely cannot assume that the activity of reducing the possible senses of the word to one is done before an eventual semantic tagger is in place. Therefore, both *flatness* and *certainty of the observations* represent a problem to be resolved.

Having a lexical hierarchy H associated to the semantic dictionary D', in absence of information the only way is to give a weight to all the active senses (as done in (Resnik, 1997) where a study of word lexical preferences is done). If a word activates n nodes in the hierarchy H each node will cumulate a $1/n$ weight in the classification function whenever encountered as training instance. For the problem we are addressing here, this model seems to disperse too much observations due to the dimension of the feature space that represents all the nodes of the hierarchy H.

We propose to use a subset of the hierarchy that we call *semantic fingerprint* subset. The *semantic fingerprint* of a word should represent all its active senses with respect to this cut of the hierarchy. Then given a hierarchy H underlying a semantic dictionary D' and a subset of nodes SF retained as a useful level of generalisations the *semantic fingerprint* of a word w, i.e. $SF(w)$, is the subset of SF activated by the word w, i.e.:

$$SF(w) = \{s \in SF | s \text{ generalises } s' \text{ and } s' \in senses(w)\}$$

where $senses(w)$ are all the senses activated by the word w in the considered hierarchy H. The set SF represents the semantic tag catalogue of the resource D', i.e. $SF = T'$.

The feature spaces we want to consider should then integrate the word and this semantic fingerprint. Two approaches are possible: a boolean and a weighted activation. The first approach tries to use the semantic fingerprint information and it is a viable solution for many ML algorithms. The second one tries to capture the relative importance between highly unambiguous and polysemous words in the training phase. Given W as the set of all the words of the dictionary and a $S_i = [0, 1]$ real interval for each element s_i in the semantic fingerprint SF, the resulting feature space is:

$$W \times S_1 \times ... \times S_n$$

where n is the cardinality of SF. The boolean model is a subcase of this as it uses only the extremes of each G_i interval. A word w instance i activating a semantic fingerprint $SF(w)$ will then have two possible representations in the feature space. The boolean activation scheme foresees w as first element and 1 for each S_i whose corresponding s_i is in $SF(w)$ and 0 for the others. The weighted activation scheme will have w as first element and $1/|SF(w)|$ for each S_i whose corresponding s_i is in $SF(w)$ and 0 for the others.

One important issue is to understand which is the most relevant semantic fingerprint. This requires to adopt different external lexical resources and different levels of generalisation, i.e. different D' and different SF within the chosen D'.

4. Probabilistic classifiers

We tested the usability of the semantic fingerprint in a probabilistic framework in order to take also profit of the weighted model. As the target is to define the classification function (1), we tried with two different stochastic estimators: a modified maximum likelihood model that takes into account the *uncertainity* of the observations and a maximum entropy model. The sample space over which probabilities have to be estimated is then the following:

$$T \times W \times S_1 \times ... \times S_n$$

where T is the set of all the semantic classes.

For the purpose of the description of the probability estimation, for each class $t \in T$ we define the function:

$$t(i) = \begin{cases} 1 \text{ if } t \text{ is the class of the instance } i \\ 0 \text{ otherwise} \end{cases}$$

and for each $s \in SF$ the function:

$$s(i) = \begin{cases} v \text{ if } v \text{ is the value of the related feature S in } i \\ 0 \text{ otherwise} \end{cases}$$

4.1. Using the Maximum Likelihood estimation in a ”back-off” approach

For this first estimation method, the probabilistic classifier is approximated with:

$$Tagger(i) \approx argmax_{t \in T} \hat{P}(t|i)$$

This latter is estimated as $\hat{P}(t|i) = max_{s \in i} P(t|w, s)$ where w is the word in i while s is one of the generalisation of w in $SF(w)$.

The estimation is then done with the following back-off model that considers the word association with the class

Test Set	*Sem. Fingerprint*	*MaxLik*	*MaxLik weighted*	*MaxEnt weighted*
Light	w	0.7748	0.7748	0.8068
	w + synset	0.7853	0.7866	0.8201
	w + BC	0.8685	0.8698	0.8673
	w + TM	0.8282	0.8527	0.8496
	w + LDOCE	0.8282	0.8201	0.8335
Hard	w	0.6830	0.6830	0.5852
	w + synset	0.6317	0.6114	0.6568
	w + BC	0.7337	0.7371	0.7342
	w + TM	0.6998	0.7002	0.7182
	w + LDOCE	0.6643	0.6608	0.6914

Table1. Experimental results

more reliable then the generalisations of the word in the semantic fingerprint:

$$P(t|w,s) = \begin{cases} \hat{P}(t|w) \text{ if } w \text{ is a seen word} \\ \hat{P}(t|s) \text{ otherwise} \end{cases}$$

The probabilities are then estimated with the maximum likelihood model as follows. Having a set of training examples Tr, the estimated probability $\hat{P}(t|w)$ is straightforwardly obtainable as:

$$\hat{P}(t|w) = \frac{counts_{Tr}(t,w)}{counts_{Tr}(w)}$$

On the other hand, the probability for the generalisation in the semantic fingerprint is estimated as:

$$\hat{P}(t|s) = \frac{\sum_{i \in Tr} t(i)s(i)}{\sum_{i \in Tr} s(i)}$$

It is worth noticing that the estimators are correctly defined for both the boolean and the weighted scheme.

4.2. Using the Maximum Entropy approach

In the Maximum Entropy model, observable features of instances are called *feature functions*. These are functions that fire in given conditions and allow the detection of some given preconditions (see (Jelinek, 1998)). Given the pair of glasses on the instance space that we have called feature-value vector, an equivalent representation can be found in terms of feature functions. The binary feature function related to the configuation (v, c) has the following form:

$$F\_v\_c(class, i) = \begin{cases} 1 \text{ if } class = c \wedge f_i = v \\ 0 \text{ otherwise} \end{cases}$$

The equivalence between a feature vector and a set of feature functions is thought in terms of representative power. If F is the i-th feature in the feature-value space, in order to represent it we will need $|F| \cdot |C|$ *feature functions* if all the configurations (v, c) with $v \in F$ and $c \in C$ are admissible. It is worth noticing that the set of feature functions can be reduced if some of these configurations are not admissible, i.e. for a given class c the feature F will never assume the value v.

If the space I is observed in the feature-value model, $F_1 \times ... \times F_n$, an equivalent (from the point of view of the expressive power) representation of this model in the ME approach will require $n \cdot |F| \cdot |C|$ feature functions.

5. Experimental Evaluation

These experiments are built to investigate if the semi-supervised approach presented in Sec. 2. is a viable solution for producing semantic taggers and if the notion of semantic fingerprint is somehow useful. Moreover, a second problem is to demonstrate that an external resource is preferable to a self-referring approach. Finally, within the chosen external semantic resource it is necessary to understand which is the more profitable cut of the hierarchy among all the possible ones.

The experiments are carried out using the annotated corpus produced in (Guthrie et al., 2004) where the target is to produce a semantic tagger able to tag with LDOCE categories. In line with what done in (Guthrie et al., 2004), we prepared two different experimental set-ups:

- a *light* test whose words kept apart in the $ToTag$ set are 194 highly ambiguous words
- an *hard* test representing the fully unsupervised model where $Train$ are all the unambiguous words of the dictionary and $ToTag$ are all the ambiguous ones

In the *light* test set the $training$ and $testing$ instances for the classification models have been obtained in the following way: the overall corpus C has been divided randomly in two parts C_1 and C_2. All the instances C_{ToTag} of the words of $ToTag$ in C_1 have been collected. The training instances Tr are then $Tr = C_1 - C_{ToTag}$ while all the testing instances Ts are $Ts = C_2 \cup C_{ToTag}$. On the other hand, in the *hard* test set, $Train$ is the portion of the dictionary that contains the unambiguous words while $ToTag$ is the set of all the ambiguous words. The Tr set is represented by all the instances in C of $Train$ words and Ts gathers all the instances in C of the $ToTag$ words.

The external semantic resource used in the experiments is WordNet and we tried three different semantic fingerprints for the nouns: (1) the synset level, no generalisation is applied and words activate their synsets; (2) the *basic concept* level, a set of WordNet synsets considered in the inter-lingual interface of EuroWordNet (Vossen, 1998); (3) the WordNet topmosts. In Table 1 these semantic fingerprints are respectively called $synset$, BC, and TM.

Two control experiments have been also carried out: one in absence of any semantic fingerprint and the second with a self-referring semantic fingerprint. Table 1 reports the results. It is possible to observe that in the case of the light experiment any use of semantic fingerprint gives a positive gain with respect to the experiment without any generalisation. Moreover, using the generalisation of an external resource is more positive than using a self-referred semantic fingerprint. It is worth noticing that the best semantic fingerprint seems to be based on the EuroWordNet base concepts. The second set of experiments on the hard test provides even more evidence on this relevant observation.

6. Conclusion

In this paper we proposed a way to use an external semantic resource in the process of semantic tagging. This has been integrated in the semantic classifiers using the notion of semantic fingerprint. With the experimental results we demonstrated that use of the semantic fingerprint helps in classifying ”unseen” words, i.e. words whose behaviour has not been manually tagged. The use of an external resource based on a more fine-grained dictionary seems to be a good solution to speed up the production of both general and domain specific semantic taggers.

7. References

Dahlgren, Kathleen G., 1988. *Naive Semantics for Natural Language Understanding*. Boston: Kluwer Academic Publishers.

Gale, William, Kennet Church, and David Yarowsky, 1992. One sense per discourse. In *Proceedings of the Speech and Natural Language Workshop*. San Francisco.

Guthrie, Louise, Roberto Basili, Fabio Massimo Zanzotto, Kalina Bontcheva, Hamish Cunningham, Marco Cammisa, Jerry Cheng-Chieh Liu, Jia Cui, Cassia Farria Martin, David Guthrie, Kristiyan Haralambiev, Martin Holub, Klaus Machery, and Fredrick Jelinek, 2004. Large scale experiments for semantic labeling of noun phrases in raw text. In *Proceedings of the Language, Resources and Evaluation LREC 2004 Conference*. Lisbon, Portugal, fortcoming.

Ide, Nancy and Jean Veronis, 1998. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–35.

Jelinek, Fredrerik, 1998. *Statistical Methods for Speech Recognition*. Cambrige, Massachussetts, USA: The MIT Press Massachusetts Institue of Technology.

Madhu, Swaminathan and Dean Lytle, 1965. A figure of merit technique for the resolution of non-grammatical ambiguity. *Mechanical Translation*, 8(2):9–13.

Miller, George A., 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Oswald, Victor A. Jr. and Richard H. Lawson, 1953. An idioglossary for mechanical translation. *Modern Language Forum*, 38(3/4):1–11.

Quinlan, J.R., 1993. *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Resnik, P., 1997. Selectional preference and sense disambiguation. In *Proceedings of ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How?, Washington, April 4-5, 1997.*.

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Vossen, P., 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Witten, Ian H. and Eibe Frank, 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Chicago, IL: Morgan Kaufmann.

Extended Semantic Tagging for Entity Extraction

Narjès Boufaden*, **Yoshua Bengio**†, **Guy Lapalme***

*Laboratoire RALI - Université de Montral
Québec, Canada
{boufaden,lapalme}@iro.umontreal.ca
†Laboratoire LISA - Université de Montral
Québec, Canada
bengioy@iro.umontreal.ca

Abstract

We present results of a statistical method we developped for the detection of what we define as generalized named entities from manually transcribed conversations. This work is part of an ongoing project for an information extraction system in the field of maritime Search And Rescue (SAR). Our purpose is to automatically detect relevant words and annotate them with concepts from a SAR ontology. Our approach combines similarity score vectors and topical information. Similarity vectors are generated using a SAR ontology and the Wordsmyth dictionary-thesaurus. Evaluation is carried out by comparing the output of the system with key answers of predefined extraction templates. Results on speech transcriptions are comparable to those on written texts in MUC7.

1. Introduction

We present a semantic labeling approach for the identification of what we call generalized named entities (GNE) from transcribed conversations. The GNE are selected types of entities same as named entities, however they are not restricted to noun phrases; they can be verbs or adjectives.

The extended semantic tagger is part of a general framework for IE pattern discovery from conversations. Targeted text is transcribed conversational speech which is more complex than transcriptions of Broadcast news (Chincor and al., 1998).In particular, the structural complexity of transcribed conversations such as turn-takings make relevant information scattered through several utterances. Speech disfluencies such as repairs and omissions alter utterances structure and increase the number of ways a given relation may be expressed. Hence, collecting relatively complete set of IE patterns from speech corpora become an even more difficult task than from texts.

Our purpose is to learn IE patterns based on GNE. In particular, we focus on the identification of generalized named entities. The extended semantic tagger is based on a statistical model which combines similarity scores and topical information. Similarity scores help identifying word groups likely to convey information related to the domain, whereas topics help distinguishing GNE from word groups which are of no particular interest.

In section 2., we present the issue of IE pattern discovery for transcribed conversations. Our approach is described in section 3. and the extended semantic tagger and its components are described in section 4. The case study in section 5. shows the results of generalized named entity extraction from transcriptions of telephone conversations in the particular domain of maritime Search and Rescue (SAR). We conclude with some proposals for further improvements.

2. IE from transcribed speech

IE is about seeking instances of class of events and relations and extracting their arguments. Despite the maturity of the information extraction (IE) tasks for written texts, IE from transcribed speech is currently restricted to the named entity task (Chincor and al., 1998). IE systems developed for well written texts use patterns based on “subject-verb-object” relation that match the sentence structure. However, whereas this is possible for well written texts where relevant event classes are expressed in a relatively easily recognizable grammatical forms, this is not the case for spontaneous speech. Two necessary hypothesis for syntax driven learning approaches are violated when processing spontaneous conversations: grammatically and locality of informations.

IE from transcribed conversational speech is a two-dimensional problem. The syntactic dimension involves the problem of disfluencies. Edited words, omissions and interruptions are examples of disfluencies that alter the utterance structure causing a significant decrease of performance in part-of-speech tagging and parsing (Charniak and Johnson, 2001). Furthermore, altering the syntactic structure of utterances make syntactic driven learning of extraction patterns difficult if not impossible.

The pragmatic dimension deals with the fact that speech and particularly conversational speech is a highly contextualized activity. Turn-takings, interruptions and overlappings, for example, result in the scattering of relevant information across a series of utterances. Tasks that require shallow or deep understanding of utterances, such as IE, must take into account a larger context than individual utterances.

3. IE pattern discovery approach

There has been considerable work on the supervised learning and *quasi* unsupervised learning of IE patterns. Supervised learning approaches use corpora which have been manually annotated to indicate the information to be extracted (Califf and Mooney, 1999; Soderland, 1999). Quasi unsupervised approaches rely heavily on syntactic information such as "subject-verb-object" relations and on a minimum annotated data; usually named entities to bootstrap the learning process (Riloff, 1998; Yangarber and al., 2000).

As far as we know, very little concern has been given to IE patterns discovery from speech corpora actually limited to Broadcast news. Most of the work has been done on texts and was first introduced to address the problem of portability of IE systems to different application domains. The reasons for this limitation are related to the structural complexity of speech. Two problems arise when learning IE patterns from transcribed conversations. Relevant information can be conveyed through successive turn-takings resulting in scattered informations and disfluencies introduce noise in data. Accordingly, patterns are not observed on utterances but on larger contexts which ensure the completeness and coherence of the conveyed informations; in this case the context is a topic segment.

The approach we present is based on supervised learning from automatically annotated transcribed conversations. Basically, we annotate GNE with domain-specific semantic labels and learn predicate-argument relations that describe IE patterns.

The IE pattern discovery process is divided into four steps. The first one is a pre-analysis of the transcribed conversations. It includes shallow parsing to detect noun groups, verbs and adjectives. The second stage is the topic segmentation and labeling. Topic segments are used as extraction units because they are larger contexts that should ensure complete "predicate-arguments" relations. The topic label represent the word context and is used to distinguish relevant entities from words of no particular interest. The third stage is the extraction of GNE. It includes a process for the recognition of known GNE and another one for the Out-Of-Vocabulary (OOV) GNE. The last stage, the IE pattern learning, is a markov model which takes as input GNE recognized in each topic segment. Figure 3. shows the different modules needed for the IE pattern learning process. In this paper, we only present the third stage which is the extraction of GNE. We tackle the problem of semantic labeling of OOV GNE (section 4.). The IE pattern learning module is left for future work and the others components are described in this section.

3.1. Domain knowledge

In IE, domain knowledge has generally been encoded in gazetteers for the named entity extraction task or in ontologies to allow inferences to generate more complex facts. In our approach, we encoded the domain knowledge in an ontology for two reasons:

- Ontologies define explicit hierarchical relations such as IS-A or PART-OF relations that can be used to generalize word classes and reduce their number to enhance the IE pattern learning process.
- They provide an interpretation or grounding of word senses, so that word sense disambiguation problem can be reduced.

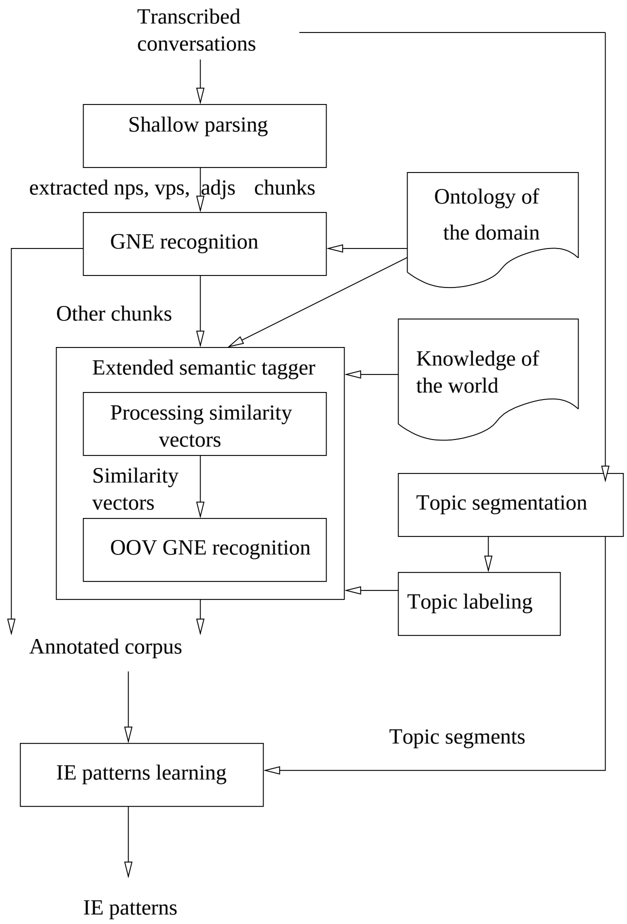

Figure1. Stages of the IE pattern discovery approach

3.2. Knowledge of the world

Dictionaries or lexicons such as WordNet are used to bridge the gap between entities from the corpus which are not described in the ontology of the domain and known entities. Thesaurus in combination with similarity measures have been previously used to enrich ontologies (Stevenson, 2002). In our approach, we used the dictionary-thesaurus Wordsmyth[1] and a similarity measure based on the overlap coefficient to assess the closeness of a word to the domain vocabulary. Figure 2 is an example of a Wordsmyth entry for the word "wonder".

3.3. Shallow parsing

Candidates to be tagged are noun groups *np*, verbs *vp* an adjectives *adj*. For this purpose, we used the Brill transformational tagger (Brill, 1992) and the CASS partial parser of Steven Abney (Abney, 1994) to parse the conversations. However, because of the disfluencies encountered in the conversations, many errors occurred when parsing large constructions. So, we reduced the set of grammatical rules used by CASS to cover only minimal chunks and discard large constructions such as VP → H=VX O=NOM? ADV* or noun phrases NP → NP CONJ NP.

[1] URL http://www.wordsmyth.net/.

```
ENT:      wonder
SYL:      won-der
PRO:      wuhn dEr
POS:      intransitive verb
INF:      wondered, wondering, wonders
DEF:      1. to experience a sensation of
                   admiration or amazement
                  (often fol. by at):
EXA:      She wondered at his bravery in combat.
SYN:      marvel
SIM:      gape, stare, gawk
DEF:      2. to be curious or skeptical
              about something:
EXA:      I wonder about his truthfulness.
SYN:      speculate (1)
SIM:      deliberate, ponder, think,
            reflect, puzzle, conjecture
...
```

Figure2. Description of the dictionnaire-thesaurus Wordsmyth entry for the verb "wonder". This verb express a request for equipement and is tagged as an instance fo the concept STATUS (8-O Figure 4). Acronymes ENT, SYL, PRO, POS, INF, DEF, EXA, SYN, SIM refers to the entry, syllable, pronunciation, part of speech, flexions, textual definitions, example, synonyms an similar words.

3.4. Generalized named entity recognition

This task, like the named entity extraction task, annotates words that are instances of the ontology. Basically, for every chunk, we look for the first match with a concept instance. The match is based on the word and its part-of-speech. When a match succeeds, the semantic tag assigned is the concept of the instance matched. Then, the semantic tag of the head is propagated to the whole chunk as shown in Figure 3.

Matching step 1: ... SN: black thicker $\underbrace{\textbf{\textit{fog}}}_{\text{WEATHER-TYPE}}$ ...

$\leftarrow$ Propagation

Propagation step: ... SN: $\underbrace{\textbf{\textit{black thicker fog}}}_{\text{WEATHER-TYPE}}$ ...

Figure3. Output of the named concept extraction process. The semantic tag of the head "fog" is propagated to the whole chunk

In (Boufaden, 2003), we show that the described approach achieves a recall score of 85,3% and a precision score of 94,8%.

3.5. Topic segmentation and labeling

The extraction unit we used is the topic segment which is composed of consecutive utterances conveying, in general, at most one piece of information that could be used to fill in a template slot. For this purpose, we developed a topic segmentation system based on a multi-knowledge source modeled by a hidden Markov model. (Boufaden and al., 2001) showed that by using linguistic features modeled by a Hidden Markov Model, it is possible to detect about 67% of topics boundaries.

The topic labeling system has not yet been developed but we are planning to develop it to fully automatically generate word context as described in our approach.

4. Extended semantic tagging

Our approach is based on psycholinguistic evidence. It has been shown that, when communicating intentions, speakers select words carefully in order to make the intention recognizable (Levelt, 1993). So, if we consider topics as indicators of communicative intentions, we can assume that given a relevant topic, words with high similarity scores are likely to convey relevant information. Hence, the relevance of a word in a specific domain can be translated into a function of word similarity to domain ontology concepts and the concepts frequency given the topic where the word appears.

In practical terms, the extended semantic tagger is a normalized product of two experts. The first expert is a similarity based model $P(C_t = k|w_t)$ that generates similarity probabilities of concepts to words from similarity scores. Whereas, the second expert is a topic based model $P(C_t = k|T_t)$ that generates concept probabilities given a topic. The product of experts is given by the equation 1.

$$P(C_t = k|w_t, T_t) = \frac{P(C_t = k|w_t)^{\beta_1} P(C_t = k|T_t)^{\beta_2}}{\Sigma_{l=1}^{K} P(C_t = l|w_t)^{\beta_1} P(C_t = l|T_t)^{\beta_2}} \quad (1)$$

$$and$$

$$C^* = \operatorname*{argmax}_{C_t} P(C_t|w_t, T_t), P(C^*|w_t, T_t) > \delta \quad (2)$$

k is one from the K concepts of the domain ontology or an Out Of Vocabulary concept (OOV), $P(C_t = k|w_t)$ is the probability that concept k is observed given the word w_t and $P(C_t = k|T_t)$ is the probability that concept k is observed given a topic T_t.β_1 and β_2 are parameters of the model

Since we are looking for GNE rather than doing only semantic tagging, we empirically determine a threshold to distinguish word groups representing GNE from non relevant words as shown in equation 2.

4.1. The similarity based model

The similarity based model generates a vector of similarity scores for each word. It uses a domain ontology and the Wordsmyth dictionary-thesaurus to determine the similarity score between a word and every concept of the domain ontology. They are computed using textual definitions of words as described in Lesk's approach (Lesk, 1996). Technical details of the algorithm used to generate similarity score vectors are described in (Boufaden, 2003). Basically, the similarity score is based on the overlap coefficient similarity measure (Manning and Schutze, 2001). It counts the number of lemmatized content words in common between the textual definition of the word and the concept. In these experiments, we do not address the word sense disambiguation problem and each similarity score is replaced by the mean of similarity scores of every word sense. We also assume conditional independence between a word and a concept $P(C_k|w(l), w) = P(C_k|w(l))$ where $w(l)$ is a word sense of w. In addition, we assume that word senses $w(l)$ are equally probable given a word w (Equation 3).

$$P(w(l)|w) = \frac{1}{|S(w)|} \quad (3)$$

Where $S(w)$ contains the different word senses of w provided by the Wordsmyth dictionary-thesaurus

Hence, $P(C_k|w)$ is given by:

$$P(C_k|w) = \Sigma_{w(l)\in S(w)} P(C_k|w(l))P(w(l)|w) \quad (4)$$

Where $P(C_k|w(l))$ is calculated from similarity scores between the concept C_k given a word sense $w(l)$ of w and $P(w(l)|w)$ is the relative frequency of the word sense $w(l)$ given w from Wordsmyth.

To process $P(C_k|w)$ we added an Out Of Vocabulary (OOV) concept for all the words that have null similarity scores for all SAR concepts. The probabilities are then generated from similarity scores by using a discounting method (Manning and Schutze, 2001).

4.2. The topic based model

The topic based model identifies the distribution of concepts given specific topics related to the domain. Basically, for every *event template* (MUC, 1998) we define a topic label. Then, each conversation is divided manually into topic segments and each topic segment is manually labeled with one of the defined topic labels or with the label *other-topic*. Concepts are classified according to equation 5.

$$P(C_t|T_t) = \alpha P_0(C_t) + (1-\alpha)P_1(C_t|T_t) \quad (5)$$

C_T are the ontology concepts, T_T topics. α is the smoothing parameter. $P_0(C_t)$ is the relative frequency and $P_1(C_t|T_t)$ is the relative frequency given a topic.

5. Case study: IE from manually transcribed SAR conversations

Our aim is to implement an information extraction system in the domain of Search And Rescue (SAR) from transcribed conversations. The conversations are mostly informative dialogs, where two speakers (a caller `C` and an operator `O`) discuss the conditions and circumstances related to a SAR mission. The conversations are either (1) incident reports, such as reporting missing airplanes or overdue boats, (2) SAR mission plans, such as requesting a SAR airplane or coast guard ships for a mission, (3) debriefings, in which case the results of the SAR mission are communicated. Figure 4 is an excerpt of such conversations. We can see that parts of some utterances were replaced by the word "INAUDIBLE" to indicate segments that have not been transcribed. In the overall corpus, such segments are found in 10% of the utterances. Besides, more than half of the corpus utterances have disfluencies such as repetitions (Ha, do, is there, is there . . .) , omissions and interruptions (we've been, _ actually had a . . .). There are about 3% transcription errors (such as `flowing` instead of `blowing` in 21-O Figure 4) which mostly occur with relevant words.

The words shown over braces in Figure 4 are the GNE to be extracted. These are, for example, the incident, its location, SAR resources needed for the mission and weather conditions. We can see the role of the topic in distinguishing entities from non relevant words. For example, in utterance 7-C the word "land" is an entity that refers to the STATUS of an airplane[2] having trouble, whereas in utterance 42-O it is of no particular interest.

5.1. SAR ontology

We built a SAR ontology using manuals provided by the National Search and Rescue Secretariat (SAR Manual, 2000) and from a sampling of 10 conversations. The ontology is composed of a sampling of key answers of predefined IE template fields such as "radar search", "diving" for means of detection, "drifting", "overdue" for incidents and "wind", "rain", "fog" for weather conditions. All were grouped into 24 semantic classes and organized in IS-A and PART-OF hierarchies. The overall ontology has a maximal depth of three. Each class represents a SAR concept and they are all used to classify entities. For each instance from the ontology we associated a list of synonyms and similar words along with their textual definitions, all extracted from Wordsmyth. Synonyms and similar words were added to increase the effectiveness of the similarity measure used.

5.2. Experiments and Results

Experiments were conducted on 4570 words that were manually annotated with SAR concepts and topic labels. 25.3% of these words are GNE. The training corpus represents 65% of the 64 manually transcribed conversations. Relevant topic segments[3] have an average length of 3 utterances. Evaluation is carried out by comparing the output of the system with key answers of predefined extraction templates. A threshold $\delta = 0.35$ was determined empirically. It means that only words that have $P(C_t|w_t, T_t) > 0.35$ are considered as GNE. Table 1 shows the precision and recall of the extended semantic tagger and the similarity based model. For the topic based model we proceed to the evaluation of the classification error. All the modules were tested on manually segmented conversations.

The major result is an assessement of the feasibility of the GNE extraction task. Our system achieves an F-score[4] of 86% which is not as good as F-scores of NEE from transcribed speech around 93% (Miller and al., 1999). However, Broadcast News are well written texts read by a speaker and can not be considered as spontaneous speech. On the other hand, our texts are spontaneous conversations with disfluencies that significantly decrease the part-of-speech tagging performance which results in increasing the semantic labeling error. Besides, since named entities are a subset of the generalized named entities we consider

[2] The airplane is actually considered as a missing object

[3] Relevant topics are topic segments that are not labeled with the 'other-topic' tag.

[4] F-score used is $F = \frac{(\beta+1)P.R}{\beta^2.P+R}$ and $\beta = 0.5$

. . .

——————— INCIDENT——————————-

7-C: On the way to go, he had to $\underbrace{\textbf{land}}_{\text{STATUS}}$ in $\underbrace{\textbf{emergency}}_{\text{INCIDENT}}$ in the $\underbrace{\textbf{South East Coast of Newfoundland}}_{\text{LOCATION}}$.

. . .

——————— SEARCH UNIT——————————-

12-O: They did $\underbrace{\textbf{a radar search}}_{\text{TASK}}$ for us in $\underbrace{\textbf{the 3 other surfaces}}_{\text{LOCATION}}$.

13-C: Hum, hum.

——————— SEARCH UNIT——————————-

18-O: And I am $\underbrace{\textbf{wondering}}_{\text{STATUS}}$ about the possibility of $\underbrace{\textbf{outputting}}_{\text{STATUS}}$ an $\underbrace{\textbf{Aurora}}_{\text{SAR-AIRCRAFT}}$ in there for $\underbrace{\textbf{radar search}}_{\text{TASK}}$.

. . .

——————— MISSION——————————-

21-O: They got $\underbrace{\textbf{a South East}}_{\text{DIRECTION}}$ to be $\underbrace{\textbf{flowing}}_{\text{STATUS}}$ there and it's just $\underbrace{\textbf{gonna}}_{\text{STATUS}}$ be $\underbrace{\textbf{black thicker fog}}_{\text{WEATHER}}$ the whole, $\underbrace{\textbf{whole South Coast}}_{\text{LOCATION}}$.

22-C:OK.

——————— OTHER-TOPIC——————————-

42-O: Now, the question he had was is there some place for a small helicopter to land there,
if he was to get something else or somebody else to take him in there ?

. . .

——————— SEARCH UNIT——————————-

56-C: Ha, they $\underbrace{\textbf{should go}}_{\text{STATUS}}$ $\underbrace{\textbf{to get going}}_{\text{STATUS}}$ at $\underbrace{\textbf{first light}}_{\text{TIME}}$.

. . .

Figure4. An excerpt of a conversation reporting an emergency landing and a request for an SAR airplane (Aurora). Numbers are utterances position in the conversation. The words in bold are extracted GNE. The tag below each bold chunk is an SAR concept from the ontology which we want to identify. Lines are boundaries of topics which were added manually (MISSION, INCIDENT, SEARCH UNIT, OTHER-TOPIC)

T1:INCIDENT

1 Initial alert	`emergency landing`
3 Location	`South East Coast of Newfoundland`
4 Date	
5 Missing object	`airplane`
7 Weather conditions	`WEATHER1`

T2:WEATHER CONDITIONS

1 Id	`WEATHER1`
2 Condition	`black thicker fog`
3 Wind direction	`South East`
4 Wind speed	
5 Visibility	

Figure5. Two filled templates from the conversation in Figure 4: the event template "INCIDENT" and the object template "WEATHER CONDITIONS".

our results comparable to those on NEE from Broadcast News.

	$P(C_t\|T_t)$	$P(C_t\|w_t)$	$P(C_t\|T_t, w_t)$
Precision	48.8%	61.0%	76.8%
Recall		55.2%	56.7%

Table1. $P(C_t|T_t)$ is topic-based model, $P(C_t|w_t)$ the similarity-based model and $P(C_t|T_t, w_t)$ the combined model. Results of the combined model are obtained with a threshold of $\delta = 0.35$. To compare the similarity based model with the combined model we tested word groups with $P(C_t|w_t) > 0.005$.

As well, results show the effectiveness of the combined model over the similarity based model $P(C_t|w_t)$. Despite the poor performance of the topic based model $P(C_t|T_t)$, it improves the detection of OOV GNE by 25.9%.

6. Conclusion

Named entity extraction (NEE) is an important stage for text based IE systems because it's a relatively easy task that has proved to be helpful for IE pattern learning. However, because of the structural complexity of transcribed speech, we believe that moving beyond named entities to identify GNE would be more helpful for the IE pattern discovery task applied to conversations.

In this paper, we experiment on the recognition of GNE related to a particular domain. The extended semantic tagger used is a stochastic model which combine similarity scores and topical information to generate semantic labels drawn from a domain ontology we designed. It is part of an ongoing work that aim to develop an IE pattern discovery method that learns predicate-arguments relations from a corpus annotated with domain-specific semantic tags.

Results of the experiments are not as good as those of related works in named entities extraction or on shallow semantic parsing (Gildea and Jurafsky, 2002). However, we believe that IE pattern based on domain-specific semantic tags is a way to get around the structural complexity of conversations.

The system being at a preliminary stage, there is room for further improvements including better smoothing in the generation of $P(C_k|w_t)$ from similarity scores. For the case study, we have worked on manually segmented conversations with manually annotated topic labels. But, we are planning to develop a system to automatically label topic segments as generated by the system described in (Boufaden and al., 2001). The last step in our project is to learn a set of IE patterns to validate our approach.

7. References

S. Abney. 1994. Partial parsing. Tutorial given at ANLP.

E. Brill. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

MUC 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7).* Morgan Kauffman.

Boufaden, N., G. Lapalme, and Y. Bengio, 2001. Topic Segmentation : A First Stage to Dialog-based Information Extraction. In *Natural Language Processing Pacific Rim Symposium, NLPRS'01*.

Boufaden, N. An ontology-based semantic tagger for ie system. In *ACL Student Workshop*, pages 7–14, Sapporo, Japon, Juillet 2003.

Chincor, N., Robinson P., and Brown E., 1998. HUB-4 Named Entity Task Definition Version 4.8. Technical report.

Califf, E. M. and Mooney R., 1999. Relational Learning of Pattern Match Rules for Information Extraction. In *16th National Conference on Artificial Intelligence AAA-I 99*.

Charniak, E. and Johnson M., 2001. Edit Detection and Parsing for Transcribed Speech. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*.

Gildea, D. and Jurafsky D., 2002. *Automatic Labeling of Semantic Roles*, volume 28(3) of *Computational Linguistics*. pages 245–288.

Grishman, R., 1997. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, SCIE-97*, volume 1299 of *Lecture Notes on Computer Sience*, chapter two. Springer Verlag, pages 10–27.

Lesk, M., 1996. Automatic Sense Disambiguation:How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the SIGDOC Conference*.

Stevenson, M., 2002. Combining Disambiguation Techniques to Enrich an Ontology. In *15th Conference on Artificial Intelligence EACI-02 workshop on "Machine Learning and natural Processing for Ontology Engineering*. lyon, France.

Manning, C. D. and Schütze, H., 2001. *Foundations of Statistical Natural Language Processing*, chapter Word Sense Disambiguiation. The MIT Press Cambridge, Massachussets London England, pages 294–303.

Miller, D., Schwartz D.R., Weischedel, R., and Stone, R.. Named entity extraction from broadcast news. In *Proceedings of DARPA Broadcast News Workshop*.

Search And Rescue Manual, 2000. Fisheries and Oceans Canada, Canadian Coast Guard, Search and Rescue, 2000. *SAR Seamanship Reference Manual*, Canadian Government Publishing, Public Works and Government Services Canada edition, November. ISBN 0-660-18352-8.

Yangarber, R., Grishman R., Tapanainen P., and Hutunen S., 2000. Automatic Acquisition od Domain Knowledge for Information Extraction. In *18th COLING Conference*. Germany.

Riloff, E., 1998. Automatically generating extraction patterns from untagged tex. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*.

Soderland, S. , 1999. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34:233–272.

Levelt, W.J.M., 1993. *Speaking: From Intention to Articulation*. MIT Press.