Semi-automatic Building Method for a Multidimensional Affect Dictionary for a New Language

Guillaume Pitel, Gregory Grefenstette

CEA-LIST
18, rte du Panorama, BP6 Fontenay-Aux-Roses F-92265 France
{guillaume.pitel, gregory.grefenstette}@cea.fr

Abstract

Detecting the tone or emotive content of a text message is increasingly important in many natural language processing applications. Examples of such applications are rating new books or movies or products, judging the mood of a customer e-mail and routing it accordingly, measuring reputation that a person or a product has in the blogosphere. While for the English language there exists a number of affect, emotive, opinion, or affect computer-usable lexicons for automatically processing text, other languages rarely possess these primary resources. Here we present a semi-automatic technique for quickly building a multidimensional affect lexicon for a new language. Most of the work consists of defining 44 paired affect directions (e.g. love-hate, courage-fear, ...) and choosing a small number of seed words for each dimension. From this initial investment, we show how a first pass affect lexicon can be created for new language, using a SVM classifier trained on a feature space produced from Latent Semantic Analysis over a large corpus in the new language. We evaluate the accuracy of placing newly found emotive words in one or more of the defined semantic dimensions. We illustrate this technique by creating an affect lexicon for French, but the techniques can be applied to any language found on the Web and for which a large quantity of text exists.

Introduction

There is an emotional dimension to many words. Some words evoke pleasant feelings and sensations and others express anger, frustration, dislike. More and more, research in computer processing of text recognizes the interest in being able to recognize this emotional level. For example, see (Abbasi and Chen, 2007). Most automated text processing systems identify this level by referring to a lexicon of affect-bearing words, such as the General Inquirer. Few other languages possess such linguistic resources. For this reason we have developed a method for bootstrapping a large-scale lexical resource for a new language. Our method involves defining semantic axes, choosing a number of seed words for each axis, and creating a number of emotion-evoking text segments to gather from the Web additional candidate words for inclusion in the affect lexicon. We then show how these candidate words are automatically classified into the semantic axes using the seed words with a classification method based on LSA and SVM. We evaluate the method by comparing the classification accuracy against a hand-built gold standard. Our techniques allows the rapid creation of a new affect lexicon for any language appearing on the Web.

1. Related Work

Given an existing affect lexicon, a number of techniques have been proposed for extending it, especially in terms of binary positive and negative polarity. Hatzivassiloglou and McKeown (1997) showed how the pattern X and Y could be used to automatically classify adjectives as having positive or negative orientation. Co-occurring words were considered to have the same polarity, and the bigger class of words was considered as having negative polarity. They achieved 92% accuracy over a set of 236 adjectives. Wiebe (2000) used a seed set of "subjective" adjectives and a thesaurus generation method (Hindle, 1990) to find more subjective adjectives. Turney and Littman (2003) developed the SO-PMI approach, using sets of known positively charged paradigm words (good, nice, excellent, positive, fortunate, correct, superior) and negative paradigm words (bad, nasty, poor, negative, unfortunate, wrong, inferior) to the classify the valence of new words, achieving 98.2% accuracy with the 334 most frequent adjectives from the (Hatzivassiloglou and McKeown, 1997) test set. Grefenstette et al. (2006) described a similar method for classifying new words into an existing, richer set of affect classes. We extend this line of work here to unresourced languages. The problem of multiclass classification has been recently evaluated on 6 classes (Anger, Disgust, Fear, Joy, Sadness, Surprise) for multiword news headlines classification (Strapparava and Mihalcea, 2007), with middling results.

2. Manually Built Resources

2.1. Defining Semantic Dimensions of Affect

Osgood et al. (1975) defined three value axes: *Positive-Negative*, *Strong-Weak*, *Active-Passive*, but hundreds of more specific axes of human values[1] can be defined. For our new language, we started from the 43 paired axes detailed in (Grefenstette et al., 2006), translating them by hand into French as shown in table 2.1. (adding an additional dimension *Admiration-Denigration*). This initial investment took three hours for one person.

2.2. Creating seed words for each dimension

For each chosen dimension, we then selected two to four seed words which we thought best captured our intuitive

[1]The Humanity Quest Web site lists more than 500 different human values, similar to our affect classes. See `http://web.archive.org/web/20031118174947/http://humanityquest.com/`.

Axis#	**Positive class**	**Negative class**	Axis#	**Positive class**	**Negative class**
01	Avantage (51)	Désavantage (28)	23	Facilitation (92)	Obstruction (83)
02	Amour (65)	Haine (44)	24	Bienfait (54)	Crime (200)
03	Entente (27)	Opposition (95)	25	Joie (130)	Tristesse (121)
04	Fidelité (28)	Traîtrise (36)	26	Bon sens (70)	Absurde (68)
05	Attraction (25)	Répulsion (18)	27	Santé (27)	Maladie (80)
06	Moralité (18)	Immoralité (31)	28	Responsabilité (21)	Irresponsabilité (12)
07	Clarté (21)	Confusion (140)	29	Honnêteté (49)	Malhonnêteté (60)
08	Protection (39)	Nuisance (42)	30	Raison (44)	Folie (88)
09	Confort (24)	Irritation (51)	31	Humilité (25)	Fierté (54)
10	Franchise (33)	Sournoiserie (38)	32	Sécurité (33)	Insécurité (24)
11	Coopération (34)	Conflit (90)	33	Amusement (65)	Horreur (74)
12	Paix (44)	Violence (39)	34	Altruisme (43)	Egoïsme (16)
13	Courage (28)	Lâcheté (48)	35	Innocence (16)	Culpabilité (27)
14	Persuasion (56)	Obligation (45)	36	Sensibilité (34)	Insensibilité (47)
15	Création (38)	Destruction (139)	37	Intelligence (55)	Stupidité (95)
16	Plaisir (77)	Douleur (63)	38	Force (55)	Faiblesse (95)
17	Désir (77)	Evitement (46)	39	Justice (22)	Injustice (37)
18	Louanges (49)	Injures (88)	40	Succès (29)	Echec (39)
19	Energie (57)	Fatigue (100)	41	Vie (36)	Mort (77)
20	Prévisibilité (29)	Surprise (69)	42	Supériorité (35)	Infériorité (44)
21	Excitation (178)	Ennui (102)	43	Abondance (68)	Manque (36)
22	Promesse (34)	Avertissement (54)	44	Admiration (75)	Dénigrement (90)

Table 1: The 44 affect axes chosen for our classification experiment. Each axis has a positive and a negative pole. In parentheses is the number of words manually affected to each axis pole.

notion of the dimension (often including the dimension label). For example for the dimension *Avantage* (advantage) we chose the **noun**, the **adjective** and the **verb** forms of the concept: *avantage, avantageux, avantager*. For *Désavantage* (disadvantage), we chose the different forms of two close synonyms: *désavantage, désavantager, désavantagée, défavoriser,* and *défavorisée*. Over the 88 dimensions, this second effort generated a list (L1) of 229 seed words. In a third step, we manually extended the list in order to reach an average 10 words per class, generating a list (L2) containing 881 words for the 88 classes. This process took about a full day for one of the authors, a native speaker of French.

2.3. Creating a gold standard for evaluation.

For subsequent evaluation purposes, we also produced a gold standard (L3) by first expanding the initial seed list (L1) using a synonyms dictionary [2] (Ploux and Victorri, 1998), and then manually deleting words that the human annotator felt intuitively did not belong to the class being built. During its construction, a few words not proposed by the synonym dictionary were occasionally added to the gold standard (L3) by the annotator. The only criterion for inclusion of a gold standard word in a dimension was: does this word significantly evoke the corresponding sentimental dimension for the native language annotator. The gold standard now contains 4980 word-to-class relations (3513 distinct words, a word can belong to more than one class), and was built in about 2 weeks. The construction of the gold standard [3] is not part of the technique for creating an affect lexicon, but was created only to evaluate the technique. Additional work to add centrality and intensity information to the word-to-class relations in underway.

[2] Available online at http://elsap1.unicaen.fr/dicosyn.html

[3] This gold standard might more appropriately be called a silver standard, since it was not built with exhaustiveness as a goal.

3. Classifying affect words along their dimensions

This first round of words was found by filtering synonyms of L2 words, leaving a collection of emotive words. The next step consists of automatically assigning these words to their appropriate semantic dimensions among the 44 we have defined. For instance, *désagrément* (*annoyance,unpleasantness*), might be assigned to the axes of *Avantage/Désavantage*,*Protection/Nuisance, Confort/Irritation, Plaisir/Douleur* and *Santé/Maladie*.

To make this assignment, we use the L1 and L2 seeds for each paired positive-negative dimension, and providing us with 44 sets of about five (for L1) or twenty (for L2) words each dimension.

3.1. Classifying with SL-dLSA+SVM

Using these sets of words, L1 and L2, we want to automatically assign new words from the L3 set in our 44 classes. In a first experiment, we evaluate the classification power of a combination of Latent Semantic Analysis (Deerwester et al., 1990) and Support Vector Machines, which have been successfully used, independently, in sentiment analysis research (Pang et al., 2002; Turney and Littman, 2002). We call this method Semantic Likeliness from diversified LSA and SVM, because we use different LSA spaces as input features for SVM. In our approach, LSA is used for its ability to reduce the number of dimensions, because SVM cannot handle too many dimensions, and we want to be able to take information from diversified cooccurrence

matrices, using different window sizes. A variety of information is provided by cooccurrence matrices derived from different windows of words: we observed that short and directed windows tend to provide highly semantic information, while bigger, symmetric ones tend to provide more thematic or pragmatic relations.

We created a total of forty-two 300-dimension semantic spaces for our language using Latent Semantic Analysis [4] from 40 million words in the French EuroParl Corpus (Koehn, 2005), varying the word windows in forty-two ways, viz., for each size in the set $\delta \in [1..10, 15, 20, 25, 30]$, we considered the windows $[0, +\delta]$, $[-\delta, +\delta]$, $[-\delta, 0]$. Each of the words was then associated with a concatenation of its LSA vectors from these spaces, producing vectors with 12600 dimensions (raw cooccurence matrices would have totalized some 5.3 million dimensions). The corpus was also prepared in the following way: it was first POS-tagged using TreeTagger (Schmidt, 1994), then transformed so that each term was a concatenation of its POS-tag and its lemma (for instance "mice" becomes "NNmouse"), and finally filtered to keep only nouns, verbs, adjectives and adverbs.

Using these 12600-dimensions LSA vectors and the LibSVM (Chang and Lin, 2001) package[5], we trained a 44 class SVM classifier[6]. Applying this classifier to our gold standard L3 (3513 words minus the 881 words already appearing in L2 or L1, leaving 2632 words that were the set tested in this and all other experiments), we obtain the results in table 2 and 3, showing the precision of classification for all unseen L3 words among 44 classes.

As expected, adding more words for training improved the classification results. The improvement seems impressive but L1 had only three to five seed words for each dimension.

An example of relatively successful classification is given in table 4 for the word "désagrément" (annoyance, unpleasantness). This example clearly illustrates that the L3 gold standard is far from exhaustive, since at least two classes might be judged acceptable by a native speake but were not in the manual classification of the word. Moreover, the class 9 can be considered as a subset of the class 16 (they could even be merged), and is probably a better choice than the gold standard class 16 for this particular word.

An example of unsuccessful classification is shown in table 5 for the word "disgrâce" (disgrace, disfavour). In this example, the classifier fails to capture the fact that a disgrace cause disadvantage, which was the rationale behind the manual assignement of the word to the class 01. However, a posteriori analyze shows that the classifier correctly capture the fact that being in disgrace means being denigrated, and that this is very close to being slandered.

[4]Using the freely available latent semantic analysis tool Infomap-NLP, see `http://infomap.stanford.edu/`.

[5]LibSVM support vector machine software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[6]Given the high dimensionnality of our model, we tried and reduce the number of dimensions using the *fselect* tool from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#4`, without any difference in the results

#Classes	***Prec.***	***Rec.***	***F***	***At least 1***
1	.16	.09	.12	.16
2	.12	.14	.13	.22
3	.10	.18	.13	.28
4	.09	.23	.13	.33
5	.08	.26	.12	.37
6	.08	.28	.12	.40
7	.07	.32	.11	.44
8	.07	.35	.12	.48
9	.07	.38	.12	.51
10	.07	.41	.12	.54

Table 2: Scores of the SL-dLSA+SVM 44 class classifier trained on **L1**, for number of responses taken into account N = 1..10 for each word to classify. 54% of gold standard words receive a proper classification among the first ten classes proposed by the classifier. The best f-score is 0.13.
A baseline based on a random assignment would have a f-score of 0.025 for $\#Classes = 1$.
#Classes is the number of classes taken into account from the classifier, ***Prec.*** is the precision of the classification, ***Rec.*** is the recall, ***F*** is the F-score ($\frac{2*Prec*Rec}{Prec+Rec}$) of the classification, and ***At least 1*** is the ratio of the number of classifications where at least one class proposed among the N classes is right.

#Classes	***Prec.***	***Rec.***	***F***	***At least 1***
1	.33	.20	.25	.33
2	.25	.29	.27	.41
3	.21	.36	.27	.48
4	.18	.42	.25	.53
5	.15	.45	.23	.57
6	.14	.49	.22	.60
7	.13	.52	.21	.63
8	.12	.55	.20	.65
9	.11	.58	.18	.68
10	.10	.60	.17	.70

Table 3: Scores of the SL-dLSA+SVM 44 class classifier trained on **L2**, for number of responses taken into account N = 1..10. 70% of gold standard words receive a proper classification among the first ten classes proposed by the classifier. The best f-score is 0.27. This shows that using more seed words for each class improves the classification.

3.2. Classifying with SL-PMI measure

In a second experiment, we used a measure inspired by the SO-PMI (Semantic Orientation Pointwise Mutual Information) (Turney and Littman, 2002). The original SO-PMI measure is intended to evaluate the positiveness/negativeness of a given word W. It is based on calculating the difference of the PMI of W with a set of arbitrarily chosen positive words (good, nice, excellent, ...) and negative words (bad, nasty, poor, ...).

For our purpose, we must adapt SO-PMI to a likeliness measure that will be used afterwards to choose the best candidate classes for a given word. We define the SL-PMI$_C$ (Semantic Likeliness Pointwise Mutual Information from

Class	*Score*
27 Health/Sickness	.105
01 Advantage/Disadvantage	.065
09 Comfort/Irritation	.065
07 Clarity/Confusion	.062
22 Promise/Warning	.056
36 Sensitivity/Insensitivity	.042
03 Amity/Anger	.040
16 Pleasure/Pain	.037
21 Excitement/Boredom	.035
24 Public-spiritedness/Crime	.034

Table 4: Classification of the word "**désagrément**" using SL-dLSA+SVM with L2, $\#Classes = 10$. Based on the gold standard L3, gold standard classes (in bold) for this word were 01 (Advantage/Disadvantage) and 16 (Pleasure/Pain). Other a posteriori acceptable classes numbers are in italic.

Class	*Score*
44 Admiration/Denigration	.144
18 Praise/Slander	.101
37 Intelligence/Stupidity	.082
13 Courage/Fear	.071
21 Excitement/Boredom	.062
16 Pleasure/Pain	.035
24 Public-spiritedness/Crime	.029
06 Morality/Immorality	.028
33 Humor/Horror	.028
23 Facilitation/Prevention	.021

Table 5: Classification of the word "**disgrâce**" using SL-dLSA+SVM with L2, $\#Classes = 10$. Based on the gold standard L3, the manually assigned class for this word (01 Advantage/Disadvantage) is absent. A posteriori acceptable classes numbers are in italic.

Information Retrieval for class C) to be :

$$\text{SL-PMI}_C(w) = \frac{1}{|C|} \sum_{c \in C} \log_2 \frac{\epsilon + H_\delta(w,c)^2}{\epsilon + H_\delta(w,*)H_\delta(c,*)}$$

Where $H_\delta(w_1, w_2)$ is the number of cooccurrences of words w_1 and w_2 in a δ words window.
Since the number of queries required for the evaluation of this method is more than 2 millions, we found it impractical to use a web search engine. Instead, we used the French SemanticMap: a resource of our own (Grefenstette, 2007), currently under construction, in which we collect the results of syntactic analysis of web pages, as well as window cooccurrence information for sizes 5, 10 and 20. At the moment of the experiment, the SemanticMap contained data from analyzing 2 million French web pages.
Applying a simple classifier based on this measure (with $\delta = 10$) to our gold standard L3 (3513 words minus the 881 words already appearing in L2 or L1), we obtain the results in table 6 and 7, showing the precision of classification for all unseen L3 words among 44 classes.
Again, the addition of training words improves the classification score, from a f-score of 0.14 to a f-score of 0.17.

#Classes	*Prec.*	*Rec.*	*F*	*At least 1*
1	.14	.11	.12	.14
2	.12	.17	.14	.22
3	.10	.22	.14	.29
4	.09	.26	.13	.33
5	.08	.30	.13	.38
6	.08	.33	.13	.43
7	.08	.37	.13	.47
8	.07	.39	.12	.51
9	.07	.42	.12	.54
10	.07	.44	.12	.56

Table 6: Scores of the SL-PMI 44 classes classifier trained on **L1**, for number of responses taken into account N = 1..10. 56% of gold standard words receive a proper classification among the first ten classes proposed by the classifier. The best f-score is 0.14.

#Classes	*Prec.*	*Rec.*	*F*	*At least 1*
1	.18	.14	.15	.18
2	.15	.21	.17	.27
3	.13	.27	.17	.35
4	.11	.32	.17	.41
5	.10	.35	.16	.45
6	.10	.38	.15	.49
7	.09	.41	.15	.53
8	.08	.43	.14	.56
9	.08	.45	.13	.58
10	.07	.47	.12	.60

Table 7: Scores of the SL-PMI 44 classes classifier trained on **L2**, for number of responses taken into account N = 1..10. 60% of gold standard words receive a proper classification among the first ten classes proposed by the classifier. The best f-score is 0.17.

The results by this method are not as good as in table 7, but the technique requies less effort, since no semantic spaces are built.

3.3. Classifying with SL-LSA measure

In a third experiment, we used a measure inherited from the SO-LSA measure (Turney and Littman, 2002). As for the SO-PMI, the original SO-LSA measure is intended to evaluate the positiveness/negativeness of a given word W. Instead of using the difference of the pointwise mutual information, it makes use of the semantic distance in a LSA semantic space.
We adapted the SO-LSA to produce a likeliness measure and we define the SL-LSA$_C$ (Semantic Likeliness from LSA for class C) to be :

$$\text{SL-LSA}_C(w) = \frac{1}{|C|} \sum_{c \in C} \cos\big(LSA_\delta(w), LSA_\delta(c)\big)$$

Where $LSA_\delta(w)$ is the vector representing word w in a LSA space built with a δ words window.
Applying a simple classifier based on this measure (with $\delta = 30$, $\delta = 10$, $\delta = 5$ and $\delta = 2$) to our gold standard L3, we obtain the results in tables 8 and 9, showing the

precision of classification for all unseen L3 words among 44 classes.

#Classes	*SL-LSA*$_{30}$	*SL-LSA*$_{10}$	*SL-LSA*$_5$	*SL-LSA*$_2$
1	.10	.11	.11	.10
2	.12	.13	.13	.11
3	.11	.12	.12	.12
4	.11	.12	.12	.11
5	.11	.11	.12	.11
6	.10	.11	.11	.11

Table 8: F-Scores for the SL-LSA$_{30}$, SL-LSA$_{10}$, SL-LSA$_5$ and SL-LSA$_2$ 44 classes classifiers trained on **L1**, for number of responses taken into account N = 1..6. The best F-score is 0.13.

#Classes	*SL-LSA*$_{30}$	*SL-LSA*$_{10}$	*SL-LSA*$_5$	*SL-LSA*$_2$
1	.13	.13	.14	.13
2	.14	.14	.15	.15
3	.14	.14	.14	.14
4	.13	.14	.14	.14
5	.13	.13	.13	.13
6	.12	.13	.13	.12

Table 9: F-Scores for the SL-LSA$_{30}$, SL-LSA$_{10}$, SL-LSA$_5$ and SL-LSA$_2$ 44 classes classifiers trained on **L2**, for number of responses taken into account N = 1..6. The best f-score is 0.15.

Using more training words again generally significantly improves the classification score. A non-significant improvement is observed for the best f-scores of SL-LSA$_{10}$.

4. Discussion

In figure 1 are shown the compared F-scores of the various classifiers trained on L1, together with the baseline from a random assignment of words to classes. The performances of our methods are very similar, with f-scores ranging from 0.11 to 0.15 for number of classes in [1..3]. Between the three kinds of classifiers (SL-dLSA+SVM, SL-PMI, SL-LSA$_*$), most of the differences for $\#Classes = 1$ are significant, but some such as SL-PMI/SL-LSA$_{5/10}$, are not (see significance values in table 10). In this presentation, SL-PMI is slightly better than other methods, followed closely by SL-dLSA+SVM, which almost attains the performance of SL-PMI for number of classes greater than 4. The performance of the SL-LSA$_*$ culminates at $\#Classes = 2$, then decreases regularly as the number of classes taken into account increases.

Using the larger seed sets L2 as the training base, the f-scores of the classifiers are those presented in figure 2. In this setting, two groups of classifiers are distinguished: SL-PMI and SL-LSA$_*$ are almost equivalent (except for $\#Classes = 1$ for which SL-PMI is slightly better), while SL-dLsA+SVM largely outperforms other classifiers, with a top result for $\#Classes = 2$. Among the three main kinds of classifiers, all the differences for $\#Classes = 1$ are significant, (see significance values in table 11). For LSA classifiers, the only significant difference lies between SL-LSA$_2$ and SL-LSA$_5$.

When comparing L2/L1 performance ratios in figure 3, we see which techniques profit most from larger seed sets. SL-dLSA+SVM rises above the other classifiers at about 2 before decreasing, the LSA family of classifiers keeps a constant ration around 1.15, and the SL-PMI classifiers quickly drops to the no improvement ratio of 1 (at $\#Classes = 2$) and even sinks in the worsening zone for $\#Classes > 5$. All the differences between L1 and L2 settings are significant with $p < 0.01$.

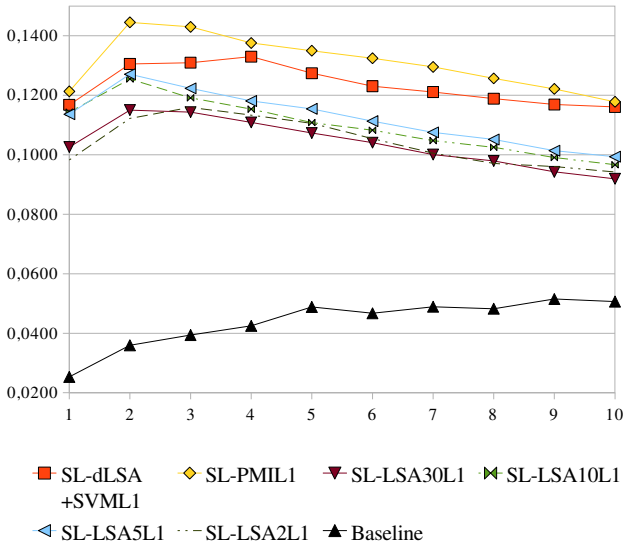

Figure 1: F-scores of the classification methods using L1 as the training data, for 1..10 best classes taken from the classifiers.

$p <$	SL-PMI^{L1}	SL-LSA$_{30}^{L1}$	SL-LSA$_{10}^{L1}$	SL-LSA$_5^{L1}$	SL-LSA$_2^{L1}$
SL-dLSA+SVM^{L1}	.01	.01	.01	.01	.01
SL-PMI^{L1}		.01	-	-	.01
SL-LSA$_{30}^{L1}$			.05	-	-
SL-LSA$_{10}^{L1}$				-	.01
SL-LSA$_5^{L1}$					.01

Table 10: Significances of differences between our settings for number of classes = 1.

5. Perspectives

5.1. Impact of the diversification of word windows

Since we did not evaluate the SVM classifier on simple LSA feature spaces, it is not easy to decide whether the use of several different LSA spaces is at the origin of the good results of the SL-dLSA+SVM when more training data is

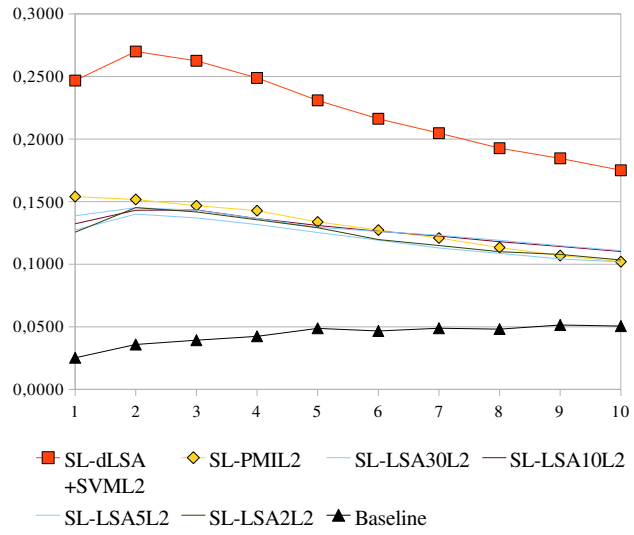

Figure 2: F-scores of the classification methods using L2 as the training data, for 1..10 best classes taken from the classifiers.

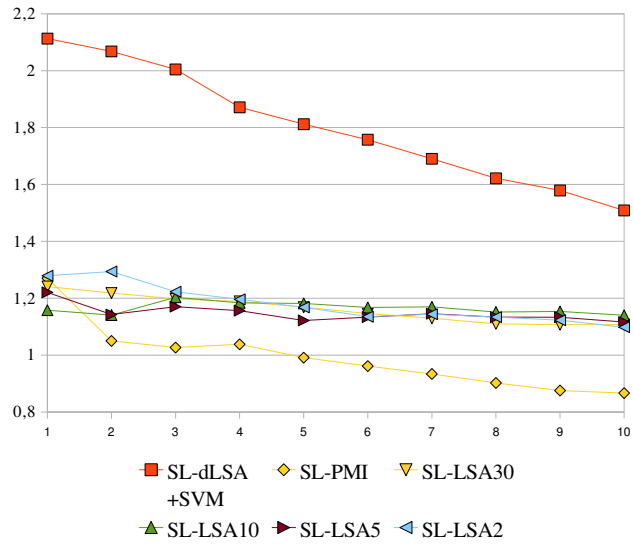

Figure 3: Improvement ratios between L2 and L1 F-scores. A ratio of 1 means no improvement.

available. The very similar results of the SL-LSA$_*$ family of classifiers could lead us to believe that their results only differ slightly. However, as shown by the table 12 the κ agreements between the four SL-LSA$_*$ variants are very low (they would be ranked as fair agreement in the traditional agreement scale).

Still, it is possible that this high disagreement only occurs when the classifiers produce wrong results. The agreements of the classifiers for their "correct" answers subset of classification is shown in figure 13. Again, the agreement is relatively low given the similarity of results from the classifiers. As a consequence, if it were possible to select the correct answers from both SL-LSA$_{30}^{L2}$ and SL-LSA$_2^{L2}$, their f-score would raise from 0.13 to 0.19.

$p <$	SL-PMI^{L2}	SL-LSA$_{30}^{L2}$	SL-LSA$_{10}^{L2}$	SL-LSA$_5^{L2}$	SL-LSA$_2^{L2}$
SL-dLSA+SVM^{L2}	.01	.01	.01	.01	.01
SL-PMI^{L2}		.01	.01	.05	.01
SL-LSA$_{30}^{L2}$			-	-	-
SL-LSA$_{10}^{L2}$				-	-
SL-LSA$_5^{L2}$					.05

Table 11: Significances of precision differences between the L2 settings for number of classes = 1. '-' means that the difference is not significant. With the exception of the difference between SL-LSA$_5$ and SL-LSA$_2$, the SL-LSA$_*$ classifiers are not significantly different from each other.

κ	SL-LSA$_{10}^{L2}$	SL-LSA$_5^{L2}$	SL-LSA$_2^{L2}$
SL-LSA$_{30}^{L2}$	.34	.28	.21
SL-LSA$_{10}^{L2}$		.39	.26
SL-LSA$_5^{L2}$			.32

Table 12: Kappa agreements between SL-LSA$_*$ classifiers.

Although this is a strong clue that using several spaces built from cooccurrence matrices with different word windows actually improve the classification potential, we must test this hypothesis in various ways :

- evaluate SVM classification from various single LSA spaces,
- evaluate Maximum Entropy and Bayesian Inference learning methods on a dLSA feature space,
- try simple combination of LSA similarity measures.

5.2. Finding candidate affect words

In order to cover a larger vocabulary than L3, we need to know which words from our language, French, have an emotive affect and need to be included in the lexicon, and along which semantic axis. Since we do not possess a full semantic lexicon for affect for our language, we generated a list of candidate affect words, by searching for the follow-

κ	SL-LSA$_{10}^{L2}$	SL-LSA$_5^{L2}$	SL-LSA$_2^{L2}$
SL-LSA$_{30}^{L2}$	.50	.43	.33
SL-LSA$_{10}^{L2}$		.49	.38
SL-LSA$_5^{L2}$			.41

Table 13: Kappa agreements between SL-LSA$_*$ classifiers for subsets of classifications for which at least one of the two classifiers produced a correct answer.

ing pattern on the Web:

$$PPro + AttVb + Adv(intensity) + X$$

where PPro is a personal pronoun, AttVb is a conjugated attributive verb (be, appear, seem) and Adv(intensity) is an intensity adverb (so, very, completely) and X is the candidate word to find.

For each of these patterns (3540 in all), we recovered up to 1000 snippets from a common search engine. For example immediately after expression *je paraissais tellement (I seemed so)*, we found the following words with their frequencies: *mort (dead)* 131 times; *antipathique (unlikable)* 64; *mal (bad)* 44; *fragile (fragile)* 42; *calme (calm)* 42; ... Combining the results from all 3540 expression, we found the following words: *mal (bad), bien (good), heureux (happy), nombreux (numerous), some prepositions, loin (far), content (contented), surprise (surprised), con (jerk), probable (probable), trop (too), sûr (sure)* ... and 10,000 other words appearing 3 or more times.

What must be done now is to train a SL-dLSA+SVM classifier using L3 data, and classify the adjectives found using the pattern-based method described above, then manually check the classification result. We must also find patterns for extracting candidate verbs, nouns and adverbs.

5.3. Refactoring the gold standard

The two classification examples in figures 4 and 5 show some important facts about the gold standard and the affect axes. Not only is the gold standard not exhaustive from the points of view of the vocabulary and the word to class assignments, but there is also a clear confusion (or at least partial overlapping) between some of the classes. For instance, the difference between Advantage/Disadvantage and Facilitation/Prevention is not clear, as well as for: Comfort/Irritation - Pleasure/Pain, Admiration/Denigration - Praise/Slander, Love/Hate - Attraction/Repulsion - Desire/Avoidance, and probably some others.

This leads to the necessary redefinition of a more robust set of affect axes, as well as setting up strict rules for the gold standard construction. This is a very open problem since none of the affect resources we know of have been built on rules more precise than "give your intuitive judgment".

Conclusion

In this paper we present a technique for high dimensional classification of sentiment-evoking words, our goal being to build a sentiment lexicon organized into 44 axes such as Love-Hate, Comfort-Irritation. We describe how we manually produce minimalist (L1), small (L2) and medium (L3) versions of such a lexicon, and evaluate various classifiers on them.

The evaluated classifiers are based on classical measures in the domain of sentiment analysis (SO-PMI and SO-LSA: Semantic Orientation based on Pointwise Mutual Information / Latent Semantic Analysis), and we also evaluate a classifier based on a SVM approach using a feature set built from a number of LSA spaces using different sizes of words windows. We show that when using the minimalist lexicon as the training set, the classifiers' performances are relatively similar. However, when using the L2 lexicon, the SVM classifier largely outperforms the others. Considering the complexity of the 44 classes classification task, the performance of our classifiers, with f-scores ranging from 0.13 to 0.27, is very respectable.

Finally, we partially back our hypothesis that using many LSA spaces brings improvement by testing the classifier based on SO-LSA with various word windows sizes, and comparing the agreement between these versions. This shows that LSA spaces based on different windows sizes produce correct answers for different classifications cases, and thus that combining them can indeed improve the performance of classifiers. It is probable that these results generalize to other domains.

Acknowledgments

This research was supported by the Jean-Luc Lagardère Foundation (`http://www.fondation-jeanluclagardere.com`), and the INRIA RAPSODIS Cooperative Research Action (`http://rapsodis.loria.fr/`).

6. References

A. Abbasi and H. Chen. 2007. Affect intensity analysis of dark web forums. In *Proceedings of IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 282–288.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407.

G. Grefenstette, Y. Qu, D.A. Evans, and J.G. Shanahan. 2006. Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In Y. Qu, J. Shanahan, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 93–106. Springer.

G. Grefenstette. 2007. Conquering Language: Using NLP on a Massive Scale to Build High Dimensional Language Models from the Web. *Proc of the 8th CICLing Conference (Mexico City, Mexico, Feb. 18-24, 2007)*, pages 35–49.

V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL.*, pages 174–181, New Brunswick, NJ.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association of Computational Linguistics.*, pages 268–275, Pittsburgh, Pennsylvania.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*.

C. E. Osgood, W.H. May, and M.S. Miron. 1975. *Cross-Cultural Universals in Affective Meaning*. University of Illinois Press.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

S. Ploux and B. Victorri. 1998. Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. *Traitement Automatique des Langues*, 39(1).

H. Schmidt. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing.*, Manchester.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.

P.D. Turney and M.L. Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERB-1094, National Research Council Canada.

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

J.M. Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence.*, pages 268–275, Menlo Park, CA. AAAI Press.