Building Affective Lexicons from Specific Corpora for Automatic Sentiment Analysis

Yves Bestgen
F.R.S-FNRS, Université catholique de Louvain, PSOR/CECL
Place du Cardinal Mercier, 10 B-1348 Louvain-la-Neuve, Belgium
E-mail: yves.bestgen@psp.ucl.ac.be

Abstract

Automatic sentiment analysis in texts has attracted considerable attention in recent years. Most of the approaches developed to classify texts or sentences as positive or negative rest on a very specific kind of language resource: emotional lexicons. To build these resources, several automatic techniques have been proposed. Some of them are based on dictionaries while others use corpora. One of the main advantages of the corpora techniques is that they can build lexicons that are tailored for a specific application simply by using a specific corpus. Currently, only anecdotal observations and data from other areas of language processing plead in favour of the utility of specific corpora. This research aims to test this hypothesis. An experiment based on 702 sentences evaluated by judges shows that automatic techniques developed for estimating the valence from relatively small corpora are more efficient if the corpora used contain texts similar to the one that must be evaluated.

1. Introduction

Automatic sentiment analysis in texts, also called opinion mining, has attracted considerable attention in recent years, primarily because of its potential use in marketing study. It aims to answer questions such as 'Is the customer who sent a mail to an after-sales service particularly dissatisfied?' 'Are the opinions about a product posted on blogs positive or negative?' 'What is the image of a political party or leader in the press?'. All these questions, which relate to the way something is presented and evaluated in a text, are particularly difficult for traditional information extraction techniques (Das & Chen, 2001; Strapparava & Mihalcea, 2007; Wilks, 1997). They present, however, many applications like transmitting to the senior members of an after-sales service the mails of which the emotional tone is the most intense.

Most of the approaches developed to classify texts or sentences as positive or negative rest on unsupervised knowledge-based methods and on a very specific kind of language resource: emotional lexicons (Andreevskaia & Bergler, 2007). These lexicons contain words tagged with their affective valence (also called affective polarity or semantic orientation) that indicates whether a word conveys a positive or a negative content. To build these resources, several automatic techniques have been proposed. Some of them are based on dictionaries and lexical databases (e.g. Esuli & Sebastiani, 2006; Kamps et al., 2004; Kim & Hovy, 2004), while others use corpora (e.g. Bestgen, 2002; Hatzivassiloglou & McKeown, 1997; Sahlgren et al., 2007; Turney & Littman, 2002, 2003).

One of the main advantages of the corpora techniques is that they can build lexicons that are tailored to a specific application simply by using a specific corpus. Currently, only anecdotal observations and data from other areas of language processing plead in favour of the utility of specific corpora (Bestgen, 2002, 2006). This research aims at testing this hypothesis explicitly.

2. Determining the affective valence of words from small corpora

To my knowledge, the first researchers to propose an automatic procedure to determine the valence of words on the basis of corpora are Hatzivassiloglou and McKeown (1997). Their algorithm aims to infer the semantic orientation of adjectives on the basis of an analysis of their co-occurrences with conjunctions. The main limitation of their algorithm is that it was developed specifically for adjectives and that the question of its application to other grammatical categories has not been solved (Turney & Littman, 2003).

If several other techniques have been proposed to determine affective valence from corpora, only a few of them have been designed to work with relatively small corpora (ten million words or fewer), a necessary property for building specific affective lexicons. Two techniques that fulfil this condition are described below.

2.1 SO-LSA

The technique proposed by Turney and Littman (2003) tries to infer semantic orientation from semantic association in a corpus. It is based on the semantic proximity between a target word and fourteen benchmarks: seven with positive valence and seven with negative valence (see Table 1).

A word is considered as positive if it is closer to the positive benchmarks and further away from the negative benchmarks. Turney and Littman proposed two techniques for estimating the strength of the semantic association between words on the basis of corpora.

The first technique estimates the semantic proximity between a word and a benchmark on the basis of the frequency with which they co-occur. Its main limitation

is that it requires a very large corpus to be effective. Turney and Littman (2003) used in their analyses a corpus made up of all the English texts available on the Internet, that is to say, some 100 billion words[1]. The origin of this limitation is in the need for observing co-occurrences between the word and the benchmark to obtain an index of proximity.

Positive words	Negative words
good	*bad*
nice	*nasty*
excellent	*poor*
positive	*negative*
fortunate	*unfortunate*
correct	*wrong*
superior	*inferior*

Table 1. The seven positive and the seven negative benchmarks used by Turney and Littman (2003).

For relatively small corpora (i.e. ten million words), they recommend the use of Latent Semantic Analysis (LSA), a mathematical technique for extracting a very large 'semantic space' from large text corpora on the basis of the statistical analysis of the set of co-occurrences in a text corpus (Deerwester et al., 1990; Landauer, Foltz & Laham, 1998).

The point of departure of the analysis is a lexical table (Lebart & Salem, 1992) containing the frequencies of every word in each of the documents included in the text material, a document being a text, a paragraph or a sentence. To derive semantic relations between words from the lexical table the analysis of mere co-occurrences will not do, the major problem being that even in a large corpus most words are relatively rare. Consequently, the co-occurrences of words are even rarer. This fact makes such co-occurrences very sensitive to arbitrary variations (Burgess, Livesay & Lund, 1998; Kintsch, 2001; Rajman & Besançon, 1997). LSA resolves this problem by replacing the original frequency table with an approximation producing a kind of smoothing effect on the associations. To this end, the frequency table undergoes singular value decomposition and it is then recomposed on the basis of only a fraction of the information it contains. Thus, the thousands of words from the documents have been substituted by linear combinations or 'semantic dimensions' with respect to which the original words can be situated again. Contrary to a classical factor analysis the extracted dimensions are very numerous and non-interpretable. One could, however, compare them to semantic features describing the meaning of words (Landauer et al., 1998).

In this semantic space, the meaning of a word is represented by a vector. To determine the semantic proximity between two words, the cosine between their corresponding vectors is calculated. The more two words (or one word and a benchmark) are semantically similar, the more their vectors point in the same direction, and consequently the closer their cosine will be to one (which corresponds to coinciding vectors). A cosine of zero shows an absence of similarity, since the corresponding vectors point in orthogonal directions. The emotional valence of a word corresponds to the sum of the cosine between this word and the positive benchmarks minus the sum of the cosine between it and the negative benchmarks.

Turney and Littman evaluated the effectiveness of their technique by comparing the predicted orientation of words with that defined in the General Inquirer Lexicon (Stone et al., 1966), which contains a list of 3596 English words labelled as positive or negative. Calculated on the basis of a corpus of ten million words, PR-LSA labels 65% of the words correctly.

2.2 DI-LSA

The technique proposed independently by Bestgen (2002), DI-LSA, is very similar to the one proposed by Turney and Littman. The main difference is at the level of the benchmarks used to evaluate a word. While SO-LSA uses a few benchmarks selected a priori, DI-LSA is based on lexicons that contain several hundred words rated by judges on the pleasant-unpleasant scale This kind of lexicon was initially developed in the field of content analysis. As early as 1965, Heise proposed to constitute a valence dictionary by asking judges to rate a sample of the most frequent English words on the pleasant-unpleasant scale. Since then, lexicons for various languages have been made up (Hogenraad et al., 1995; Whissell et al., 1986). As an example, Table 2 shows the evaluation scores of several words randomly extracted from the dictionary and used in the present study (Hogenraad, Bestgen & Nysten, 1995).

Word	Valence	Word	Valence
détresse *distress*	1.4	contrôlable *controllable*	3.5
imbécile *idiotic*	1.4	outil *tool*	4.3
tristesse *sadness*	1.6	risquer *to risk*	4.5
hostilité *hostility*	2.2	entier *entirety*	4.9
impassible *impassive*	2.6	revenir *to return*	5.0
superstitieux *superstitious*	2.8	admiratif *admiring*	5.7
hâte *hastens*	3.1	doux *sweet*	6.0
ambigu *ambiguous*	3.2	sincérité *sincerity*	6.1

Table 2. Emotional valences of several words on a scale from very unpleasant (1.0) to very pleasant (7.0).

To determine the emotional valence of a word on the basis of the words with which it co-occurs in a corpus, a

[1] Turney and Littman (2003) used for this analysis the Altavista search engine.

specific whole set of benchmarks is selected from the lexicon. More precisely, the unknown valence of a word corresponds to the average valence of its thirty closer neighbours, neighbourhood being identified on the basis of the cosine in the semantic space extracted by LSA. To evaluate this index, Bestgen (2002) compared the predicted values for words with their actual values according to the dictionary and obtained correlations ranging from 0.56 to 0.70. He also showed that taking into account the thirty closer neighbours yields a better estimate than taking into account only five neighbours.

3. Experiment

This experiment aims to determine the effect on an automatic sentiment analysis task of the similarity between test materials and a corpus from which an affective lexicon is extracted. The materials for this sentence-level classification experiment (Riloff, Wiebe, & Phillips, 2005) are composed of 702 sentences [2] published in the Belgian French-language newspaper *Le Soir* in 1995.

These sentences were evaluated by ten judges. Their task was to indicate, on a seven-point scale, up to what point the contents of each sentence evoked an unpleasant, neutral or pleasant idea.

Participants read, individually and in a different random order, the 702 sentences of the corpus on a computer screen. The sentences were successively displayed just above the rating scale. Participants gave their ratings by clicking on the button corresponding to the level of pleasantness they felt. A 'validate' button enabled them to confirm their choice and to start processing the next sentence. Participants could pause at any time. The instructions specified that a break of at least one hour was to be taken around the middle of the task. On average, the participants took fifteen seconds to rate each sentence. Table 3 provides some examples of the sentences and their emotional valence.

The inter-rater agreement, computed by means of Cronbach's alpha, was very high (0.93). The average correlation between the ratings of one participant and the average ratings of all the other participants was 0.75 (the leave-one-out technique). The average correlation between the ratings provided by two participants was 0.60. A detailed presentation of the procedure used to build the materials is given in Bestgen, Fairon and Kevers (2004).

3.1 Method

The two techniques described above were used in this experiment. The fourteen SO-LSA benchmarks chosen by Turney and Littman (2003) were translated into French (*bon, gentil, excellent, positif, heureux, correct et supérieur: mauvais, méchant, médiocre, négatif, malheureux, faux et inférieur*). For DI-LSA, a French lexicon made up of 3000 words evaluated on the pleasant-unpleasant scale was used (Hogenraad et al., 1995). A minimum of thirty judges rated the words on a seven-point scale from 'very unpleasant' (1) to 'very pleasant' (7).

Valence	Sentence
2.90	Tom est papa pour la troisième fois depuis ce 18 janvier. *Tom became a father for the third time on 18 January.*
2.10	J'ai été sacré révélation de 2ème division avec un point d'avance sur Max ! *I was awarded best young player of the 2nd division with one point more than Max!*
1.00	En fait, le fameux chimiste était devenu tellement célèbre que plus personne n'avait rien à lui refuser, sauf Max fils. *In fact, the famous chemist had become so renowned that nobody could refuse him anything except his son Max.*
0.00	Marc est le seul responsable socialiste important, cité jusqu'ici à la barre des témoins. *Marc is the only important socialist leader called as a witness.*
-1.00	À l'entrée comme à la sortie de Luc , pas un seul journaliste ne s'est levé. *As Luc came in and left the room, no journalist stood up.*
-2.00	Moins chanceux, Tom dut alors défendre un déficit de 4.750.000 francs. *Less lucky, Tom then had to justify a deficit of 4.750.000 Fr.*
-3.00	Vexé, furieux, Pierre a exécuté Max puis abattu de sept balles le collègue réveillé par la première détonation. *Upset, furious, Pierre killed Max then fired seven bullets at a colleague awoken by the first detonation.*

Table 3 : Emotional valences of several sentences.

Three corpora of five million words each, varying in similarity to the test materials, were used to estimate the proximity between the words and the benchmarks:

- Soir1995 Corpus. This includes newspaper articles published in *Le Soir* during the early months of 1995: that is, the period from which the target sentences were extracted.
- Soir1997 Corpus. A comparable corpus was built from the articles published in *Le Soir* during the early months of 1997.
- Literary Corpus. A literary corpus of texts was built from novels and short stories available on the Web (mainly in the literary Web databases ABU and Frantext).

The Soir1995 corpus is most similar to the test materials. The Soir1997 corpus includes texts from the same source as the test materials, but from a later period. The Literary

[2] Each sentence was automatically modified so as to replace the name and the description of the function of every individual by a generic first name of adequate sex (Mary, John, etc.) in order to prevent the judges being influenced by their prior positive or negative opinion about these people.

corpus contains texts from a very different genre: it is the least similar to the test materials.

To be able to compare these three corpora in a fair way, the three semantic spaces were extracted, one from each corpus, according to an identical procedure adapted from Bestgen (2006). These corpora were subdivided into segments of 125 words. All the words of a segment had to come from the same text. All the segments of fewer than 125 words (articles of small size and the last incomplete segment of a text) were removed. These rules produced 40635 segments for the Literary corpus and more than 50000 for the other two corpora. In order to be able to compare corpora of different types, but of same sizes, only the first 40635 segments of the Soir1995 and Soir1997 corpora were taken into account.Singular value decomposition was realised with the program SVDPACKC (Berry, 1992), and the first 300 singular vectors were retained.

3.2 Results

The predicted valences, corresponding to the average valence of the words belonging to the sentence were compared with the judges' ratings by means of Pearson's correlation coefficients (see Table 4). Two levels of reference to measure the effectiveness of the techniques are given by previous analyses of the test materials (Bestgen et al., 2004). First, a correlation of 0.39 was obtained between the judges' ratings of the sentences and those based on the original lexicon of 3000 words, a value statistically significant ($p<0.0001$). In order to determine the effectiveness of a lexicon which takes into account all the words included in the sentences, two judges were asked to decide if each word present in the sentences, but absent from the original lexicon, was positive, neutral or negative. The correlation between the sentence ratings and that obtained on the basis of this exhaustive dictionary was 0.56.

The most important result has a bearing on the large difference in efficiency between the three corpora used to compute the word-benchmark similarities. Both techniques are far more efficient when the word's affective valence is estimated from the semantic proximities in corpora that contain texts very similar to the one from which the test materials have been extracted. Interestingly, there is little difference between the Soir1995 and the Soir1997 corpora, leading to the conclusion that it is the genre or the source of the texts that matters and not the fact that the test materials and the semantic space were extracted from the very same texts.

As regards the difference between the techniques, DI-LSA outperforms SO-LSA as well as the original lexicon. It even almost reaches the level of efficiency of the manually-expanded lexicon. If SO-LSA only weakly outperforms the lexicon approach, this performance is notable because SO-LSA is based on only fourteen benchmarks while the original lexicon includes 3000 words evaluated by numerous judges.

Corpus	SO-LSA	DI-LSA
Soir95	0.43	0.54
Soir97	0.42	0.51
Literary	0.30	0.44

Table 4: Correlations between the sentence valence as estimated by the judges' ratings and by the two techniques on the basis of the three corpora.

4. Conclusion

The experiment reported above shows that automatic techniques developed for estimating the valence of words from relatively small corpora (five million words) are more efficient if the corpora used contain texts similar to the one that must be evaluated. Obviously, the beneficial effect of using a corpus similar to the test materials would have been more strongly supported if the opposite demonstration could have been carried out: the Literary corpus should outperform the newspaper ones when the test materials are made up of sentences extracted from literary texts.

More generally, it seems that in the present experiment we are close to the maximum effectiveness of the lexical approach to evaluating sentences, since the automatic technique is nearly as effective as the traditional approach based on an exhaustive dictionary. The correlation between the predicted valences and the valences obtained from judges, however, is just higher than 0.50. If one wishes to go beyond this level of efficiency, it is probably essential to combine lexical information and more complex linguistic analyses. The 'simplistic' character of an approach based solely on the words considered individually has been strongly criticised (Bestgen, 1994; Pang et al., 2002; Polanyi & Zaenen, 2003). For example, Polanyi and Zaenen (2003) underline the need to take into account negations but also some connectors ('Although Boris is brilliant at maths, he is a horrific teacher') and the modal operators ('If Mary were a terrible person, she would be mean to her dogs'). It is, however, noteworthy that these criticisms of the lexical approach do not reject it but underline the need to supplement it. Having the most powerful lexical indices possible is a prerequisite for following this new avenue of research.

5. Acknowledgements

Yves Bestgen is a research fellow of the Belgian National Fund for Scientific Research (F.R.S.-FNRS) and a member of the Centre for English Corpus Linguistics. This work was supported by a grant (Action de Recherche concertée) from the government of the French-language community of Belgium.

6. References

Andreevskaia, A., Bergler, S. (2007). CLaC and CLaC-NB: Knowledge-based and corpus-based approaches. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval-2007),

pp. 117–-120,
Berry, M. W. (1992). Large scale singular value computation. *International Journal of Supercomputer Application*, 6, 13--49.
Bestgen, Y. (1994). Can emotional valence in stories be determined from words ? *Cognition and Emotion*, 8, 21--36.
Bestgen, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes. In *Actes du Colloque International sur la Fouille de Texte*, pp/ 81--94.
Bestgen, Y. (2006). Improving text segmentation using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings and Moore (2001). *Computational Linguistics, 32*, pp. 5--12.
Bestgen, Y., Fairon, C., Kevers, L. (2004). Un baromètre affectif effectif. *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 182--191,
Burgess C., Livesay K., Lund K., (1998). Explorations in Context Space : Words, Sentences, Discourse, *Discourse Processes*, 25, pp. 211--257.
Das S., Chen M. (2001). Yahoo! for Amazon : Opinion extraction from small talk on the web, Working Paper (under review), Décenbre 2001, Santa Clara University.
Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (1990). Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41, pp. 391--407.
Esuli, A., Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC 2006), pp. 417--422).
Hatzivassiloglou, V., McKeown, K.R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Meeting of the Association for Computational Linguistics*, pp. 174--181.
Heise, D.R. (1965). Semantic differential profiles for 1000 most frequent english words, *Psychological Monographs*, 79, pp. 1--31.
Hogenraad, R., Bestgen, Y., Nysten, J.L. (1995). Terrorist Rhetoric : Texture and Architecture, In E. Nissan & K.M. Schmidt (Eds.), *From Information to Knowledge*, pp. 48--59, Intellect.
Kamps, J., Marx, M., Mokken, R. J., De Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings of LREC 2004*,, pp. 1115--1118.
Kintsch W., (2001). Predication, *Cognitive Science*, 25, pp. 173--202.
Kim, S., Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of COLING-04, 20th International Conference on Computational Linguistic*, pp. 1367–-1373.
Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25, pp. 259--284.
Lebart, L., Salem, A. (1992). *Statistique textuelle*. Dunod.
Pang, B., Lee, L., Vaithyanathan, V. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in natural language processing*, pp. 79--86.
Polanyi, L., Zaenen, A. (2003). Shifting attitudes. In P*roceedings of Multidisciplinary approaches to discourse 2003*, pp. 61--69.
Rajman M., Besançon R., (1997). Text Mining: Natural Language Techniques and Text Mining Applications, *Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics*, Chapam & Hall.
Riloff, E., Wiebe, J., Phillips, W. (2005). Exploiting Subjectivity Classification to Improve Information Extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence* (AAAI-05).
Sahlgren, M., Karlgren, J., Eriksson, G. (2007). SICS: Valence annotation based on seeds in word space. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval-2007), pp. 296–-299,
Strapparava, C., Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval-2007).
Stone, P. J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.
Turney, P., Littman, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *Technical Report*, National Research Council Canada.
Turney, P., Littman, M. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems* (TOIS), 21, pp. 315--346
Wilks, Y. (1997). Information Extraction as a Core Language Technology. In M. T. Paziensa (Ed.) *Information Extraction,* pp. 1–-9, Springer.
Whissell C.M., Fournier M., Pelland R., Weir D., Makarec K., (1986). A dictionary of affect in language : IV. Reliability, validity, and applications ", *Perceptual and Motor Skills*, 62, pp. 875–-888.
Yu, H., Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 129–-136.