Automatic extraction of polar adjectives for the creation of polarity lexicons

Silvia VÁZQUEZ[1] Muntsa PADRÓ[1] Núria BEL[1] Julio GONZALO[2]

(1)UNIVERSITAT POMPEU FABRA, Roc Boronat, 138, Barcelona, Spain
(2) E.T.S.I. INFORMÁTICA UNED, Juan del Rosal, 16, Madrid, Spain
{silvia.vazquez, muntsa.padro, nuria.bel}@upf.edu,
julio@lsi.uned.es

ABSTRACT

Automatic creation of polarity lexicons is a crucial issue to be solved in order to reduce time and efforts in the first steps of Sentiment Analysis. In this paper we present a methodology based on linguistic cues that allows us to automatically discover, extract and label subjective adjectives that should be collected in a domain-based polarity lexicon. For this purpose, we designed a bootstrapping algorithm that, from a small set of seed polar adjectives, is capable to iteratively identify, extract and annotate positive and negative adjectives. Additionally, the method automatically creates lists of highly subjective elements that change their prior polarity even within the same domain. The algorithm proposed reached a precision of 97.5% for positive adjectives and 71.4% for negative ones in the semantic orientation identification task.

KEYWORDS: Sentiment Analysis, Opinion Mining, Polarity Lexicon, Subjectivity Detection

Proceedings of COLING 2012: Posters, pages 1271–1280,
COLING 2012, Mumbai, December 2012.

1 Introduction

In recent years, Sentiment Analysis has become one of the most important applications of Natural Language Processing. In the beginning, the discipline tried to reutilize techniques used in fields like Document Classification, Information Extraction or Question-Answering, but soon researchers realized that the typology of the texts in Sentiment Analysis was very different from those studied in these areas (Cardie, 1997), (Stoyanov, Cardie, & Wiebe, 2005). In this sense, for the summarization of subjective texts, the most important issue is to discover what is the general and predominant opinion, evaluation, emotion or speculation expressed by the author, and not the identification of the main topic of the text, the main interest of the cited areas. This task can only be done with information about the polarity of words.

Discovery and extraction of the vocabulary used to express subjectivity is crucial to start the development of any complex sentiment analysis tool. For example, knowing that an old film could be positive for some people but negative for others is very important in order to summarize the global opinion of that product. Therefore, designing algorithms that allow us to automatically build these kinds of language resources is very important.

There are three main approaches to create polarity lexicons: manual, dictionary based and corpus based. Early works in the field of Sentiment Analysis manually compiled lists of subjective words but this task was very time consuming and needed great human efforts. Some examples of this approach are The General Inquirer (Stone, Dunphy, Smith, & Ogilvie, 1966) and some of the lists of verbs annotated by Levin (Levin, 1993).

Dictionary based approach utilizes external language resources as lexicons and thesaurus which, although not collecting polarity relations, can help to increase the number of a set of opinion seeds by different methods. The majority of the works that follow this procedure make use of WordNet (Miller, 1995) to carry out this task. In the work of (Hu & Liu, 2004) the authors hypothesized that synonyms of a seed adjective have the same semantic orientation while the antonymous would have the opposite one, employing WordNet synsets to find out these relations. Lexical resources like SentiWordNet (Esuli & Sebastiani, 2005) (Baccianella, Esuli, & Sebastiani, 2010) classified polarity elements into Positive, Negative or Objective by analyzing the similarity between the glosses or definitions of the words and also by studying the relations established among them in the thesaurus. Valitutti (Valitutti, Strapparava, & Stock, 2004) tried to adapt WordNet to Sentiment Analysis purposes through the identification and subsequent annotation of all the elements having a high load of emotion or affective content.

Although the dictionary based approach achieved great results, it has two main shortcomings. On the one hand, it does not take into account the polarity changes due to different domains. As some works demonstrated (Vázquez & Bel, 2012), a great majority of the adjectives are domain dependent: they could be positive in one domain but negative or even neutral in another. On the other hand, this approach suffers from a lack of scalability since it does not take into account words not appearing in the language resources used. Actually, it falls down on the analysis of colloquial words or different kinds of slang expressions that are not collected in WordNet or any thesaurus.

Corpus based approach starts, as dictionary based one, with a manually built list of seed words but unlike it, this approach does not rely on the availability of external language resources (that for some languages could even not exist) but on linguistic cues which systematically appear in opinionated texts. The main idea behind this approach is that there are actually linguistic constraints that allow automatically identifying opinion-bearing words. One of the most early and

well-known work that followed this method was proposed by Hatzivassiloglou and McKeown (1997). This work will be commented in more detail in Section 2. Other important works based on this approach are (Kanayama & Nasukawa, 2006), (Kaji & Kitsuregawa, 2007) and (Riloff, Wiebe, & Wilson, 2003). Kanayama and Nasukama tried to expand a set of polar atoms (words and expressions) starting from an unannotated corpus and an initial lexicon. Their main assumption was that opinion words with the same prior polarity appear successively in the text, unless this context changed through an adversative expression. Kaji and Kitsuregawa addressed the polarity lexicon building from the lexico-syntactic patterns found in a large collection of documents. They achieved high precision for positive (92%) and negative (88%) elements but their recall is low. The work of Riloff et al. was not restricted to adjectives but they collected subjective nouns (they managed to learn 1000 new subjective nouns) by a bootstrapping process.

In this paper, we follow the corpus-based approach and propose a bootstrapping method to automatically and iteratively extract polar adjectives as well as their prior polarity. Additionally, this bootstrapping method permits to identify all of the polar adjectives that, exclusively depending on the context (i.e. surrounding words), can behave as positive or negative polar elements. The proposed method achieved a precision of 97.5% for positive adjectives and 71.4% for negative ones in the semantic orientation identification task and significantly increased recall to 67%.

The remainder of this paper is organized as follows. Section 2 introduces the methodology followed in our experiment, the bootstrapping process carried out and the results achieved. Section 3 details the evaluation of the bootstrapping method proposed. Finally, we present the conclusions and outline the future work.

2 Methodology

The contribution of our method to automatically identify, extract and label subjective adjectives is that we introduce a bootstrapping approach to gain coverage, and a new category of adjectives, i.e. "highly subjective adjectives", to gain precision. Our method is based, basically, on the following two works.

We based our method on the approach presented in Hatzivassiloglou & McKeown (1997) where the authors hypothesized that two adjectives joined by "and" have the same semantic orientation while two adjectives joined by "but" have the opposite one. They used this idea along with a log-linear regression model and a set of supplementary morphological rules to predict whether a pair of adjectives joined by any of these conjunctions has the same or different semantic orientation. Once pairs of adjectives are extracted, they utilized a clustering method to separate all the adjectives conjoined into two groups. The group with more elements was labeled as positive adjectives and the other as negative. This final labeling task, based on the normal frequency of positive elements, it is right if we work with a balanced corpus (with the same number of positive and negative reviews). However, in the case we worked with a corpus with more negative than positive texts, the number of negative words tended to be higher, and, therefore, the results of the tagging could be biased.

In this work, they achieved a 92% of accuracy in the classification of positive and negative adjectives.

The second work in which our research is based on is (Vázquez & Bel, 2012). This work is a case study where the authors introduced a taxonomy of polar adjectives. The results of their study showed that a great majority of polar adjectives change their prior polarity values when occurring

in different domains, that is, an adjective could be positive in a domain but negative or even irrelevant in other. For example "entertaining" is very positive in a film review, but has no sense, for instance, in a car review. Besides, the authors proposed a new type of polar adjectives, called "highly subjective adjectives", which could change their prior polarity not only among different domains but even within the same domain. For instance, a "big" car, could be positive for some customers (easy to park) but negative for others (any space inside).

To consider the existence of these "highly subjective" adjectives turned out to be very important in our experiments to gain precision. Taking into account the existence of these kinds of units in our bootstrapping process, it was possible to automatically discover not only domain dependent positive and negative adjectives but also to identify highly subjective adjectives that had caused mistakes in our final lexicon if we had not identified them.

The bootstrapping algorithm that we propose automatically extracts all of the polar adjectives joined by "y" ("and") or "pero" ("but") in a given corpus. A small set of seed adjectives as well as their corresponding prior polarity values is used for initializing the algorithm. This initial seed list was made from domain independent adjectives, therefore these elements could be used as initial list of seeds not only in the domain of cars, but also in any domain that we want to work with.

Our methodology differs from the one proposed by Hatzivassiloglou and McKeown since we hypothesized that after the first detection step, the new adjectives and their corresponding prior polarity can be iteratively reused to discover more new polar adjectives. We utilized the adjectives that were in our seed polarity lexicon as input for our algorithm to find new adjectives joined with them, identifying also the prior polarity of those. Therefore, we propose that polar adjectives and their corresponding polarity values can be automatically identified if they are found in a coordinated construction with the appropriate conjunctions and with other adjectives that were not in our seed lexicon. The process will continue until any adjective of our lexicon is not found joined with any new adjective or until there is no more conjunctive relation of this type.

Additionally, following the taxonomy of polar adjectives proposed in (Vázquez & Bel, 2012), we also automatically built lists of elements that should be treated differently in order to avoid important mistakes in the precision of automatically built polarity lexicons. As Vázquez & Bel (2012) we have worked with Spanish. However the method can be applied to any language where the conjunctive constructions work in the same manner.

Therefore, our algorithm operates on the following conditions:

- If a seed adjective is joined by "y" ("and") with an unknown adjective (that is, it is not in our seed list) and did not appear in contradictory constructions[1], we will conclude that the unknown adjective will have the same semantic orientation of the seed adjective and can be added, along with its corresponding prior polarity, to our polarity lexicon.
- If a seed adjective is joined by "pero" ("but") with an unknown adjective and did not appear in contradictory constructions, we will conclude that the unknown adjective will have the opposite semantic orientation of the seed adjective and can be added, along with its prior polarity, to our polarity lexicon.

[1] Positive adjective + and + negative adjective; negative adjective + and + positive adjective; positive adjective + but + positive adjective ; negative adjective + but + negative adjective

- If a seed adjective appears in conjunctive patterns which imply that its semantic orientation is positive but also appears in conjunctive patterns which imply that its semantic orientation is negative, the polar adjective will be added to the highly subjective adjective list.

See a diagram of the process in FIGURE 1.

2.1 Bootstrapping experiment

As explained before, the bootstrapping algorithm was meant to iteratively increase the number of polar adjectives collected for our polarity lexicon as well as to separate elements in our highly subjective adjective lists.

The experiment was carried out using a corpus of 250,000 words from car reviews. This corpus was extracted from a wider corpus (8 million of words) consisting of texts of different domains (cars, movies, mobile phones, video games and sport teams).

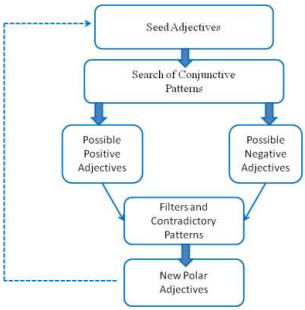

FIGURE 1 – Diagram of the bootstrapping process

All of the texts were collected from Ciao[2], a website specialized in reviews where the users write in Spanish, the language studied in this work, and where they are paid for doing this task. This last aspect guaranteed us a minimum level of correctness in all the texts, minimizing the amount of noisy text in the study.

The corpus was annotated with Part-Of-Speech tags and lemmatized using Freeling[3] POS tagger (Padró, Collado, Reese, Lloberes, & Castellón, 2010) and indexed using Corpus Query Processor (CQP)[4] (Christ, 1994) in order to facilitate the search of coordinated adjectives.

The process started by searching adjectives in the corpus occurring in a set of conjunction patterns, in order to find all the adjectives that were conjoined. 482 pairs of adjectives joined by the conjunctions "y" ("and") or "pero" ("but") were found. These pairs were the input for the identification of polarity if joined with an adjective of a known polarity; in a first step if the pair contains an adjective of the seed list, and later if containing an adjective identified and labeled by the algorithm.

[2] http://www.ciao.es/
[3] http://nlp.lsi.upc.edu/freeling/
[4] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

We started the iterative process with 28 positive and 7 negative seed adjectives. These elements were taken from the list of Bel and Vázquez (2012). Seeds were very reliable polar words that five annotators manually labeled as domain independent in a previous work. See the lists of positive and negative seeds in TABLE 1.

The procedure was iteratively repeated until no more polar adjectives were extracted, and it finished in 7 iterations.

2.2 Results

As a result of the bootstrapping process proposed in the last subsection, we increased six times the number of polar adjectives that there were in the seed polarity dictionary. We augmented the positive adjectives from 28 (seeds) to 173 and the negative ones from 7 (seeds) to 37. Crucially, we identified 13 highly subjective adjectives that indeed appeared with positive polarity in some contexts and with negative in others.

Positive Seeds[5]	alucinante, bello, bueno, chulo, cojonudo, elegante, espectacular, estupendo, excelente, excepcional, extraordinario, fantástico, genial, hermoso, impecable, impresionante, increíble, inmejorable, insuperable, lindo, magnífico, maravilloso, novedoso, perfecto, precioso, recomendable, sensacional, único
Negative Seeds[6]	terrible, pésimo, malo, horrible, feo, cutre, chungo

TABLE 1- Lists of positive and negative seeds

The growth in the number of adjectives in connection with the number of iterations is detailed in FIGURE 2.

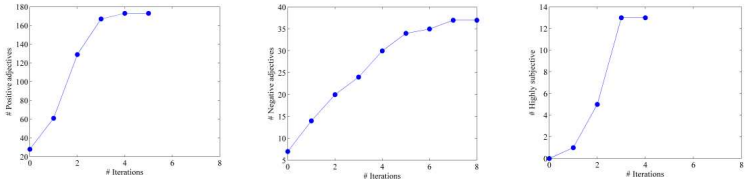

FIGURE 2 – Positive, negative and highly subjective adjectives collected and number of iterations

3 Evaluation

In this section, we report on the evaluation of the bootstrapping method proposed in the Section 3. To carry out this evaluation, we manually annotated a Gold Standard which consisted of the 12% of the whole car corpus; 200 documents in total. In each text, all the polar adjectives that should be in the final polarity lexicon were identified and labeled with their corresponding semantic orientation (positive or negative) in the particular context where they appeared. For the

[5] Amazing, beautiful, good, lovely, brilliant, elegant, spectacular, excellent, exceptional, extraordinary, fantastic, terrific, impeccable, impressive, incredible, unbeatable, pretty, superb, marvellous, original, perfect, gorgeous, sensational (some of them are synonyms so we avoided to repeat them in the translation)

[6] Terrible, dreadful, bad, horrible, ugly, shabby, dicey

annotation task we used Brat[7] (Stenetorp et al., 2012), a web-based annotation tool that allowed us to create our own labels, adapted to the experiment.

The instructions given to the annotator were the following: "If an adjective is used to describe a positive or negative speaker's evaluation, opinion, emotion or speculation of some of the objects reviewed, then this word should be in our polarity lexicon and annotated with the label that better describe it according to its semantic orientation".

It is important to note that some words that are typically used as subjective elements can also be found as objective ones. For example, "pequeño" ("small") behaves as a subjective adjective in sentences like "este coche es pequeño y aburrido" ("this car is small and boring") where we can easily understand than the writer does not like the car, since he joined the adjective "pequeño" ("small") with a negative adjective, in this case, "aburrido" ("boring"). However, if the writer was enumerating the general characteristics of the car, for example in "este coche es pequeño ya que solo tiene dos plazas, tiene 3 puertas y los vidrios tintados…" ("this car is small because it only has two seats, has three doors and dyed glasses…"), it does not imply that "pequeño" ("small") was positive or negative. In this last example, the writer performed a merely informative function, the adjective acting as an objective unit. In these cases, if the adjective had a subjective behavior, it was annotated with its corresponding tag, while if it was objective remained untagged.

The Gold Standard contained 263 words annotated as polar adjectives, being 199 of them tagged as positive and 52 of them as negative. See some examples of the annotated adjectives in TABLE 2.

It is important to note that 12 of them were identified as highly subjective elements since they were tagged as positive in some occasions and as negative in others. Some examples are "alto" ("high"), "grande" ("big") or "pequeño" ("small").

Label	Examples
Positive Adjective	afortunado, bestial, deportivo, poderoso[8]
Negative Adjective	despreciable, renqueante, molesto, prohibitivo[9]

TABLE 2 – Examples of annotation in the Gold Standard

In order to evaluate the bootstrapping process proposed in Section 2.1, we repeated the experiment only with the texts that formed the Gold Standard. We searched for all the conjunctive patterns and found 64 pairs of adjectives joined by "y" ("and") or "pero" ("but"). Therefore, we collected 64 pairs of adjectives of the total of 482 appearing in the car corpus. Then, we repeated the bootstrapping process carried out for the conjunctive pairs extracted from the car corpus, over the pairs of conjoined adjectives extracted from the Gold Standard.

Obviously, in this case, the growth in the number of adjectives collected is smaller, since we worked only with a 13% of the total pairs of adjectives joined by a conjunction. We augmented the positive adjectives from 28 (seeds) to 55 and the negative ones from 7 (seeds) to 14. In these data, we did not identify any highly subjective adjective due to the reduction of the corpus.

[7] http://brat.nlplab.org/
[8] Lucky, terrific, sports, powerful
[9] Despicable, ailing, annoying, prohibitive

The recall of the bootstrapping process proposed was calculated comparing the total number of adjectives that appeared in the conjunction pairs with the number of polar adjectives that our method was capable to extract. We identified the 67% of all the adjectives that appear in the 64 pairs of adjectives.

In order to know the precision of the method, we calculated the number of adjectives that were correctly labeled (as positive or negative) over all of the adjectives extracted by the bootstrapping process. In the Gold Standard, of all the 51 adjectives identified, 41 of them were tagged as positive, 7 of them were tagged as negative and 3 of them were extracted of these lists as highly subjective because they appeared labelled as positive or as negative depending on the context. This yields a precision for positives of 97.6% and 71.5% for negatives. See all the results in TABLE 3.

Recall	Precision for Positives	Precision for Negatives
67%	97.6%	71.5%

TABLE 3 – Recall and precision of the extraction and annotation of the polar adjectives

The results of the experiment and the data obtained with the evaluation show that our bootstrapping algorithm is able to identify and label most of the polarity adjectives contained in a corpus. The evaluation shows that our method achieves better rates of precision than other published works reported in Section 1 while maintaining recall.

Conclusions and future work

In this paper we present a bootstrapping method to automatically identify, extract and label polar adjectives, not only as positive or negative but also as highly subjective elements. Our method is based on the hypothesis that two adjectives joined by "y" ("and") have the same prior polarity and two adjectives joined by "pero" ("but") have the opposite one. Additionally, it labels as "highly subjective" all of the adjectives that can behave as positive as well as negative depending on the context. This triple classification of the polar adjectives improves the methods based on the same hypothesis and achieves a precision of 97.6% in the identification and labeling of positive elements and of 71.5% in the classification of negative ones.

Moreover, our method is capable to extract some slang polar adjectives, (for example, "cojonudo" ("insane"), "fardón" ("showy")) since it is not based on external language resources but in the real language usages of the writers. Apart from that, it is possible to reutilize the bootstrapping method because the process is simple and replicable for other domains and languages.

In future works, we will adapt the bootstrapping method proposed in order to extract and annotate polar nouns joined with the appropriate conjunctions and we also plan to study the possible extractions of polar verbs and adverbs.

Acknowledgments

This work was funded by the EU 7FP project 248064 PANACEA and the UPF-IULA PhD grant program.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, et al. (Eds.), Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10 (Vol. 0, pp. 2200–2204). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

Cardie, C. (1997). Empirical Methods in Information Extraction. AI Magazine, 18(4), 65–79. Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1322

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System, 10. Computation and Language. Retrieved from http://arxiv.org/abs/cmp-lg/9408005

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. Proceedings of the 14th ACM international conference on Information and knowledge management CIKM 05, 617–624. doi:10.1145/1099554.1099713

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. Proceedings of the 35th annual meeting on Association for Computational Linguistics, pages(2), 174–181. doi:10.3115/976909.979640

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD *'04* (p. 168). New York, New York, USA: ACM Press. doi:10.1145/1014052.1014073

Kaji, N., & Kitsuregawa, M. (2007). Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. Computational Linguistics, 43(June), 1075–1083. Retrieved from http://www.aclweb.org/anthology/D/D07/D07-1115

Kanayama, H., & Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing EMNLP 06, (July), 355–363. doi:10.3115/1610075.1610125

Levin, B. (1993). English Verb Classes and Alternations: A Preliminary Investigation. Chicago Il (p. 366). University of Chicago Press. Retrieved from http://www.amazon.com/English-Verb-Classes-Alternations-Investigation/dp/0226475336

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39–41. doi:10.1145/219717.219748

Padró, L., Collado, M., Reese, S., Lloberes, M., & Castellón, I. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools. Retrieved from http://upcommons.upc.edu/e-prints/handle/2117/7616

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In E. Riloff, J. Wiebe, & T. Wilson (Eds.), Proceedings of the seventh conference on Natural language learning at HLTNAACL 2003 (Vol. 4, pp. 25–32). Association for Computational Linguistics. doi:10.3115/1119176.1119180

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics

(pp. 102–107). Avignon: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/E12-2021

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). The General Inquirer: A Computer Approach to Content Analysis. (M. I. T. Press, Ed.)The MIT Press (Vol. 08, p. 651). MIT Press. Retrieved from http://www.webuse.umd.edu:9090/

Stoyanov, V., Cardie, C., & Wiebe, J. (2005). Multi-Perspective Question Answering Using the OpQA Corpus. Proceedings of HLT-EMNLP 2005. Retrieved from http://www.cs.cornell.edu/home/cardie/papers/hlt-emnlp05-ves.pdf

Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing Affective Lexical Resources. PsychNology Journal, 2(1), 2004.

Vázquez, S., & Bel, N. (2012). A Classification of Adjectives for Polarity Lexicons Enhancement. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/223_Paper.pdf