Learning Sentiment Lexicons in Spanish

Verónica Pérez-Rosas, Carmen Banea, Rada Mihalcea

Department of Computer Science
University of North Texas
{veronica.perezrosas, carmen.banea}@gmail.com, rada@cs.unt.edu

Abstract

In this paper we present a framework to derive sentiment lexicons in a target language by using manually or automatically annotated data available in an electronic resource rich language, such as English. We show that bridging the language gap using the multilingual sense-level aligned WordNet structure allows us to generate a high accuracy (90%) polarity lexicon comprising 1,347 entries, and a disjoint lower accuracy (74%) one encompassing 2,496 words. By using an LSA-based vectorial expansion for the generated lexicons, we are able to obtain an average F-measure of 66% in the target language. This implies that the lexicons could be used to bootstrap higher-coverage lexicons using in-language resources.

Keywords: multilingual natural language processing, multilingual subjectivity and sentiment analysis, lexicon generation

1. Introduction

Subjectivity and sentiment analysis focuses on the automatic identification of private states, such as opinions, emotions, sentiments, evaluations, beliefs, and speculations in natural language. While subjectivity classification labels text as either subjective or objective, sentiment classification adds an additional level of granularity, by further classifying subjective text as either positive, negative or neutral. A large number of text processing applications have already used techniques for automatic sentiment and subjectivity analysis, including expressive text-to-speech synthesis (Alm et al., 2005), tracking sentiment timelines in on-line forums and news (Lloyd et al., 2005; Balog et al., 2006), analysis of political debates (Thomas et al., 2006; Carvalho et al., 2011), question answering (Yu and Hatzivassiloglou, 2003), and conversation summarization (Carenini et al., 2008).

Much of the research work to date on sentiment and subjectivity analysis has been applied to English, but work on other languages is growing, including Japanese (Kobayashi et al., 2004; Suzuki et al., 2006; Takamura et al., 2006; Kanayama and Nasukawa, 2006), Chinese (Hu et al., 2005; Tsou et al., 2005; Zagibalov and Carroll, 2008), German (Kim and Hovy, 2006), and Romanian (Mihalcea et al., 2007; Banea et al., 2008b). In addition, several participants in the Chinese and Japanese Opinion Extraction tasks of NTCIR-6 (Kando et al., 2008) performed subjectivity and sentiment analysis in languages other than English.

As only 27% of Internet users speak English,[1] the construction of resources and tools for subjectivity and sentiment analysis in languages other than English is a growing need. In this paper, we propose a new method to build a subjectivity and sentiment lexicon for Spanish, which we will later employ to perform sentence level sentiment classification, as well as seek to enrich through a bootstrapping process in the target language.

[1] www.internetworldstats.com/stats.htm, Oct 11, 2011

2. Related Work

Lexicons have been widely used for sentiment and subjectivity analysis, as they represent a simple, yet effective way to build rule-based opinion classifiers. For instance, one of the most frequently used lexicons is the subjectivity and sentiment lexicon provided with the OpinionFinder distribution (Wiebe and Riloff, 2005). The lexicon was compile from manually developed resources augmented with entries learned from corpora, and it contains 6,856 unique entries that are also associated with a polarity label, indicating whether the corresponding word or phrase is positive, negative, or neutral. SentiWordNet (Esuli and Sebastiani, 2006) is a resource for opinion mining built on top of WordNet, which assigns each synset in WordNet with a score triplet (positive, negative, and objective), indicating the strength of each of these three properties for the words in the synset. The SentiWordNet annotations encompass more than 100,000 words and were automatically generated, starting with a small set of manually labeled synsets.

While there are several English lexicons for sentiment and subjectivity analysis, we are only aware of a very small number of such lexicons available for other languages. (Abdul-Mageed et al., 2011) manually compiled a list of approximately 4,000 Arabic adjectives from the newswire domain annotated for polarity. (Clematide and Klenner, 2010) extract a list of 8,000 nouns, verbs, and adjectives in German annotated for polarity and strength. Most efforts to date, though, have focused on automatic procedures of lexicon construction, such as (Kaji and Kitsuregawa, 2007) for Japanese, (Lu et al., 2010; Xu et al., 2010) for Chinese, or (Banea et al., 2008a) for Romanian. The work closest to ours is authored by (Rao and Ravichandran, 2009), who introduce a lexicon induction method that uses the WordNet graph and the relationships it entails to extend polarity classification to other words using graph based semi-supervised learning algorithms, such as mincuts, randomized mincuts, and label propagation. The latter method is the best performing one and was applied to Hindi (employing the Hindi WordNet[2]) and to French (using the OpenOffice thesaurus[3]). Our work is different in that it only explores the WordNet structure to extract parallelism across languages, and does not make use of the embedded additional relations such as hypernymy, hyponymy, meronymy, antonymy, etc., and to a limited extent synonymy. Thus while they use WordNet for within-language polarity propagation, we use it for cross-language expansion.

3. Learning Subjectivity and Sentiment Lexicons

While manually constructing a subjectivity and polarity lexicon in other languages is desirable, this process is both time and resource intensive, and thus prohibitive. A less costly approach may be importing such information from other languages with readily available electronic resources (Mihalcea et al., 2007) and then manually or automatically filtering and growing the newly acquired lexicons within each language.

In this study we seek to answer two questions. First, can we generate a reliable sentiment lexicon in a target language by using manually annotated lexicons from a source language? Second, if a manually annotated dataset is not available, could we use instead resources that have a higher coverage due to being automatically annotated for sentiment? As (Mihalcea et al., 2007; Wan, 2008) have shown that simply translating a subjectivity or polarity lexicon in a target language (in their experiments the languages are Romanian and Chinese, respectively) using a bilingual dictionary does not create a high accuracy resource due to the highly overloaded meaning of words, we seek to sidestep this issue by employing a multilingual sense aligned lexical ontology.

First, we attempt to make use of the manual annotations embedded within the Opinion Finder lexicon that are available at the word level. Since subjectivity and polarity are qualities that were shown to be most robustly expressed in English at the sense level (Wiebe and Mihalcea, 2006), we attempt to transfer the manual annotations onto the English WordNet by enforcing SentiWordNet (Esuli and Sebastiani, 2006) based constraints. The criterion we applied in order to select the corresponding sense was to match the Opinion Finder manually assigned polarity strength (i.e., strong positive and strong negative) to the sense with the highest polarity score (positive or negative) present in SentiWordNet, and then transfer the label to the English WordNet sense. This way we are able to select a polar sense with a high degree of confidence of it actually displaying a polar charge. Afterward, we draw upon the fact that the entire multilingual WordNet family uses aligned *synsets*[4] as building blocks, which allows for unequivocal sense-level mapping among languages. Thus, this method would be able to port manually annotated polarity and subjectivity information from English in any of the approximately fifty languages in which non-commercial WordNets are available[5] and thus afford what we will call a *full strength lexicon*.

Second, since the amount of manually annotated sentiment data in English is nonetheless limited, we use a second method that allows us to leverage resources that are automatically annotated for sentiment at the sense level in English. Since these resources are automatically generated, they espouse a higher coverage at the cost of lower precision, when compared to manually annotated data. We thus transfer the scores provided by a resource such as SentiWordNet by traversing the same multilingual synset-aligned WordNet structure. In the end we are able to generate a secondary *medium strength lexicon* in the target language.

To demonstrate these methods, we focus on Spanish as the language in which we seek to develop sentiment lexicons, and we employ the Spanish WordNet[6] for our experiments and evaluations. The generated lexicons are publicly available.[7]

For the first method, we start out with the single word entries available in the OpinionFinder lexicon annotated as either strong positive or strong negative. This choice is motivated by the fact that the entries are annotated at the *word level*, yet we seek to transfer these annotations at the *sense level*, and thus we need the words to exhibit a reliable sentiment content. Let us consider the word "devastation" and its annotations extracted from the OpinionFinder lexicon:

devastation - part-of-speech: noun
- type: weak subjective
- polarity: strong negative

Once we query SentiWordNet for "devastation," five synset offsets are returned, with varying positive and negative scores (see Table 1). From these we select the highest score matching the manually assigned polarity label, and thus map "devastation" to the synset 00967157 in WordNet 3.0. As the Spanish WordNet is aligned to WordNet 1.6, we locate the corresponding translation based on the immutable sense key identifier across all WordNet versions. This allows us to obtain the translation into Spanish as "devastación," and add it to our *full strength lexicon*. We thus are able to generate a lexicon containing 1,347 entries. In those rare situations where we are not able to resolve the synset alignments between different WordNet versions (such as a synset from WordNet 1.6 gaining additional granularity, or being merged with another synset in WordNet 3.0), we discard the conflicting cases.

In order to leverage the additional automatic annotations contained in SentiWordNet, in the second method we rely only upon the polarity scores that are higher than 0.5, and translate their respective synsets into Spanish, thus obtaining a *medium strength lexicon* of 2,496 entries.

[2] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[3] http://www.openoffice.org/

[4] A *synset* represents a grouping of entities (be they nouns, verbs, adverbs or adjectives) that share a distinct meaning or sense, and its members can be used interchangeably in the same context.

[5] http://www.globalwordnet.org/gwa/wordnet_table.htm

[6] http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=31&Itemid=57

[7] http://lit.csci.unt.edu/

Syn Offset	Pos	Neg	WordNet Definition
00967157	0	0.625	Plundering with excessive damage and destruction.
07335414	0	0	An event that results in total destruction.
07509827	0.25	0.5	The feeling of being confounded or overwhelmed; "her departure left him in utter devastation."
14562142	0	0	The state of being decayed or destroyed.
00217014	0	0	The termination of something by causing so much damage to it that it cannot be repaired or no longer exists.

Table 1: SentiWordNet annotations for the synsets in which the noun "devastation" appears and the corresponding WordNet definition.

4. Lexicon Evaluations

As our lexicons were compiled using English resources that are manually and / or automatically annotated for polarity, we expect them to carry strong polarity clues. In order to evaluate each lexicon's quality, we perform an evaluation using machine learning over a vector representation of the entries, and seek to discriminate between positive and negative words.
Since a raw word is unable to provide sufficient information regarding its polarity charge on its own when considering a machine learning setup, we explore the possibility of infusing semantic information to create a vector space model. Previous research in this regard (Maas et al., 2011) has suggested such a representation is able to capture both semantic information and a rich sentiment content. If we were to use context-based word unigrams as features, this would allow for a very sparse data representation, further requiring a large corpus to train a viable machine learning model. Instead, we opt to implement semantic expansion using Latent Semantic Analysis (LSA) (Dumais et al., 1988) to generate concept vectors for each of the lexicon entries. The vectors were obtained using the INFOMAP software[8] trained on an approximately 55 million words Spanish corpus, consisting of publicly available novels listed on Project Gutenberg[9] and the 2008 Spanish version of Wikipedia. We were able to obtain 44 features resulted from performing singular value decomposition on the word count matrix derived from this corpus. We then build two training sets: in the case of the *full strength lexicon*, the class is assigned based on the manual annotations, while for the *medium strength lexicon*, it is automatically ascribed to the label having the highest polarity score.
Since these evaluations were done on data annotated for polarity in English, we decided to double our evaluations by annotating 100 entries from each lexicon in Spanish, this time. This setup would further allow us to appraise how reputable an English gold standard would be in evaluating out-of-language lexicons. Annotations were performed by two native speakers of Spanish, agreement was 91% and 85% for the lexicons, respectively. Upon discussing disagreements, two gold standards emerged. These test sets were subsequently removed from our lexicons, and we train SVM classifiers on the resulted models.
The results pertaining to these evaluations are presented in Table 2.

[8] http://infomap-nlp.sourceforge.net/
[9] http://www.gutenberg.org/

Gold standard	Method	Class	P	R	F
Full strength lexicon					
EsTest1	SVM	Pos	64.6%	82.4%	72.4%
		Neg	74.3%	53.1%	61.9%
EsTest1	Manual	Pos	91.8%	88.2%	90.0%
		Neg	88.2%	91.8%	90.0%
Medium strength lexicon					
EsTest2	SVM	Pos	73.7%	54.9%	62.9%
		Neg	62.9%	79.6%	70.3%
EsTest2	Manual	Pos	85.1%	67.8%	75.4%
		Neg	64.1%	82.9%	72.3%

Table 2: Lexicons evaluations

5. Discussion

In terms of class suggestion made by the OpinionFinder lexicon when projected onto Spanish compared to the Spanish gold standard (EsTest1), the F-measure is 90% for both positive and negative labels; the manual annotations, whether assigned in Spanish or in English coincide 90% of the time. For the second lexicon (using the SentiWordNet based labels) weighed against the Spanish gold standard (EsTest2), the F-measure drops to 75.4% for the positive class and to 72.3% for the negative class, while the labels agree 74% of the time. These metrics capture the fact that the lexicon generated using the first method is, as expected, more accurate than leveraging automatically assigned scores. However, the coverage of automatically generated resources may make up for what is lost in precision (see discussion below). These two types of lexicons should be seen as complementary in allowing resource generation in a target language.
Wanting to verify whether this trend carries on to the machine learning experiments, we notice that the overall F-measure of 67.2% for the first lexicon drops to 66.6% for the second. This lower gap of 0.55%, in comparison to the drop by 16.2% experienced in the case of manual evaluations on the first and second lexicons is explained by two factors. First, the second lexicon has 85% more entries than the first, thus allowing a larger train set to provide input for the classification task. Second, the vectorial expansion using LSA allows context based co-occurrence metrics to participate in the decision process, complementing the sentiment dimension carried by the lexicon entry alone. While the class distribution for the first Spanish test set is 51 positive to 49 negative, for the second test set the distribution is more skewed towards the positive class (e.g. 64 positive to 36 negative). This explains why the lowest performing precision of 62.9% occurs for the negative class of the second lexicon. The reader should also not forget that the two lexicons represent disjoint sets.
While (Mihalcea et al., 2007; Wan, 2008) have tried automatically translating sentiment lexicons developed for English using bilingual dictionaries, the accuracies obtained

by these resources has been low. By using the WordNet structure and its main appeal of relating words based on senses both within and between languages, we are able to provide a sensible solution to sentiment lexicon export with an accuracy of approximately 90% for Spanish.

6. Conclusion

We presented a framework that generates sentiment lexicons in a target language by using manually and automatically annotated English resources. The manual annotations performed in the target language show that the first lexicon has an accuracy of 90%, since it leverages manual English annotations, while the second lexicon (which uses automatically assigned SentiWordNet scores) attains an accuracy of 74%. This demonstrates that we are able to obtain better results when using a multilingual sense aligned resource (such as the WordNet structure enriched in a number of languages) than when using multilingual dictionaries.
Furthermore, machine learning experiments using feature expansion for the extracted lexicons offer a precision higher than 62.9% for both the positive and the negative classes. This allows us to explore this venue further in future work by attempting to bootstrap the derived lexicons in the target language, as well as using them to train sentence level classifiers demonstrating a higher coverage than what could be achieved with rule-based classifiers using sentiment lexicons.

Acknowledgments

This material is based in part upon work supported by National Science Foundation award #0917170. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. References

Muhammad Abdul-Mageed, Mona T Diab, and Mohammed Korayem. 2011. Subjectivity and Sentiment Analysis of Modern Standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, number 3, pages 587–591, Portland, Oregon, US.

C. Alm, D. Roth, and R. Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, Canada.

K. Balog, G. Mishne, and M. de Rijke. 2006. Why are they excited? identifying and explaining spikes in blog mood levels. In *Proceedings of the 11th Meeting of the European Chapter of the As sociation for Computational Linguistics (EACL-2006)*.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008a. A Bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Learning Resources Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008b. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 127–135, Honolulu, Hawaii.

G. Carenini, R. Ng, and X. Zhou. 2008. Summarizing emails with conversational cohesion and subjectivity. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008)*, Columbus, Ohio.

P. Carvalho, L. Sarmento, J. Teixeira, and M. Silva. 2011. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the Association for Computational Linguistics (ACL 2011)*, Portland, OR.

Simon Clematide and Manfred Klenner. 2010. Evaluation and Extension of a Polarity Lexicon for German. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 7–13.

Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88*, pages 281–285, New York. ACM Press.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.

Yi Hu, Jianyong Duan, Xiaoming Chen, Bingzhen Pei, and Ruzhan Lu. 2005. A New Method for Sentiment Classification in Text Retrieval. *Natural Language Processing IJCNLP 2005*, 3651:1–9.

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, number June, pages 1075–1083, Prague. Computational Linguistics.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, number July, pages 355–363. Association for Computational Linguistics.

Noriko Kando, Teruko Mitamura, and Tetsuya Sakai. 2008. Introduction to the NTCIR-6 Special Issue. *ACM Transactions on Asian Language Information Processing (TALIP)*, 7(2):4:1—-4:3.

Soo-Min Kim and Eduard Hovy. 2006. Identifying and Analyzing Judgment Opinions. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-2006)*, pages 200–207, New York, New York, USA.

Nozomi Kobayashi, Kentaro Inui, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting Evaluative Expressions for Opinion Extraction. In *Proceedings of IJCNLP 2004*, number June, pages 596–605.

L. Lloyd, D. Kechagias, and S. Skiena. 2005. Lydia: A system for large-scale news analysis. In *String Processing and Information Retrieval (SPIRE 2005)*.

Bin Lu, Yan Song, Xing Zhang, and Benjamin K Y Tsou. 2010. Learning Chinese polarity lexicons by integration of graph models and morphological features. In *IInformation Retrieval Technology 6th Asia Information Retrieval Societies Conference AIRS 2010*, volume 6458/2010, pages 466–477. Springer.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 142–150, Portland, Oregon, US.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning Multilingual Subjective Language via Cross-Lingual Projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007)*, pages 976–983, Prague, Czech Republic.

Delip Rao and Deepak Ravichandran. 2009. Semi-Supervised Polarity Lexicon Induction. In *The 12th Conference of the European Chapter of the Association for Computational Linguistics*, number April, pages 675–682, Athens, Greece.

Yasuhiro Suzuki, Hiroya Takamura, and Manabu Okumura. 2006. Application of Semi-supervised Learning to Evaluative Expression Classification. In *Proceedings of the 7th International Conference on Intelligent Text*, pages 502–513.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2006. Latent Variable Models for Semantic Orientations of Phrases. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.

M. Thomas, B. Pang, and L. Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, Australia.

Benjamin K Y Tsou, Raymond W M Yuen, Oi Yee Kwong, Tom B Y Lai, and Wei Lung Wong. 2005. Polarity Classification of Celebrity Coverage in the Chinese Press. In *Proceedings of the International Conference on Intelligence Analysis*.

Xiaojun Wan. 2008. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 553–561, Honolulu.

Janyce Wiebe and Rada Mihalcea. 2006. Word Sense and Subjectivity. In *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL-2006)*, Sydney, Australia.

Janyce Wiebe and Ellen Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497, Mexico City, Mexico.

Ge Xu, Xinfan Meng, and Houfeng Wang. 2010. Build Chinese Emotion Lexicons Using A Graph-based Algorithm and Multiple Resources. *Computational Linguistics*, (August):1209–1217.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 129–136, Sapporo, Japan.

Taras Zagibalov and John A. Carroll. 2008. Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text. In *Proceedings of the Conference on Computational Linguistics (COLING 2008)*.