Cross-Linguistic Sentiment Analysis: From English to Spanish

Julian Brooke
Department of Linguistics
Simon Fraser University
Burnaby, BC, Canada
jab18@sfu.ca

Milan Tofiloski
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
mta45@sfu.ca

Maite Taboada
Department of Linguistics
Simon Fraser University
Burnaby, BC, Canada
mtaboada@sfu.ca

Abstract

We explore the adaptation of English resources and techniques for text sentiment analysis to a new language, Spanish. Our main focus is the modification of an existing English semantic orientation calculator and the building of dictionaries; however we also compare alternate approaches, including machine translation and Support Vector Machine classification. The results indicate that, although language-independent methods provide a decent baseline performance, there is also a significant cost to automation, and thus the best path to long-term improvement is through the inclusion of language-specific knowledge and resources.

1. Introduction

Sentiment analysis refers to the automatic determination of subjectivity (whether a text is objective or subjective), polarity (positive or negative) and strength (strongly or weakly positive/negative). It is a growing field of research, especially given the gains to be obtained from mining opinions available online. Approaches to sentiment analysis have tackled the problem from two different angles: a word-based or semantic approach, or a machine learning (ML) approach. The word-based approach uses dictionaries of words tagged with their *semantic orientation* (SO), and calculates sentiment by aggregating the values of those present in a text or sentence [17]. The ML approach uses collections of texts that are known to express a favorable or unfavorable opinion as training data, and learns to recognize sentiment based on those examples [13].

Our approach is semantic, and makes use of a series of dictionaries, additionally taking into account the role of negation, intensification and irrealis expressions. We believe that a semantic approach offers the advantage of taking many different aspects of a text into account.

One of the disadvantages of a semantic approach is that the resources necessary for a new domain or a new language need to be built from scratch, whereas a machine-learning approach only needs enough data to train. In this paper we show that porting to a new language, Spanish, requires only a small initial investment, while providing the opportunities for further improvement available only to semantic methods.

For comparison, we have taken three approaches to performing sentiment analysis in a new language. Our main approach involves deploying Spanish-specific resources, which we build both manually and automatically. The second approach, used in Bautin et al. [4] and Wan [18], consists of translating the texts into English, and using an existing English calculator. Finally, the third approach builds unigram Support Vector Machine classifiers from our Spanish corpora.

Our evaluation on multi-domain corpora indicates that, although translation and machine learning classification both perform reasonably well, there is a significant cost to automated translation. A language-specific SO Calculator with dictionaries built using words that actually appear in relevant texts gives the best performance, with significant potential for improvement.

2. The English SO Calculator

Our semantic orientation calculator (SO-CAL) uses five main dictionaries: four lexical dictionaries with 2,257 adjectives, 1,142 nouns, 903 verbs, and 745 adverbs, and a fifth dictionary containing 177 intensifying words and expressions. Although the vast majority of the entries are single words, our calculator also allows for multiword entries written in regular expression-like language.

The SO-carrying words in these dictionaries were taken from a variety of sources, the three largest a corpus of 400 mixed reviews from Epinions.com, a 100 text subset of the 2,000 movie reviews in the Polarity Dataset [12], and positive and negative words from the General Inquirer dictionary [15]. Each of the open-class words were given a hand-ranked SO value between 5 and -5 by a native English speaker. The numerical values were chosen to reflect both the prior polarity and strength of the word, averaged across likely interpretations. For example, the word *phenomenal* is a 5, *nicely* a 2, *disgust* a -3, and *monstrosity* a -5. The dictionary was later reviewed by a committee of three other researchers in order to minimize the subjectivity of ranking SO by hand.

SO-CAL also implements a modified version of contextual valence shifting as originally proposed by Polanyi and Zaenen [14], including negation and intensification. We have also added irrealis blocking.

Our approach to negation differs from Polanyi and Zaenen's in that negation involves a polarity shift instead of a switch: A negated adjective is shifted by a fixed amount (4) toward the origin. This means that the negation of a strongly negative word (like *terrible*) will be neutral or weakly negative (*not terrible* -5 + 4 = -1 instead of 5), while the negation of a weakly positive word like *nice* is equally negative (*not nice* 2 – 4 = -2).

International Conference RANLP 2009 - Borovets, Bulgaria, pages 50–54

The calculation of intensification is somewhat more sophisticated than simple addition and subtraction. Each expression in our intensifier dictionary is associated with a multiplier value. For instance, *very* has a value of .25, which means the SO value of any adjective modified by *very* is increased by 25%. We also included three other kinds of intensification that are common within our genre: the use of all capital letters, the use of exclamation points, and the use of discourse *but* to indicate more salient information (e.g., *…but the movie was GREAT!*).

Some markers indicate that the words appearing in a sentence might not be reliable for the purposes of sentiment analysis. We refer to these using the linguistic term *irrealis*. Irrealis markers in English include modals (*would, could*), some verbs (*expect, doubt*), and certain kinds of punctuation (questions, quotations). When SO-carrying words appear within the scope of these markers, our calculator ignores them.

Lexicon-based sentiment classifiers generally show a positive bias [10], likely the result of a human tendency to favor positive language [6]. In order to overcome this bias, we increase the final SO of any negative expression (after other modifiers have applied) by a fixed amount (currently 50%).

For initial testing, we use the 400 text Epinions corpus (50 texts in each of eight different product types), the other 1,900 texts in the Polarity Dataset (Movie), and a 2,400 text corpus of camera, printer, and stroller reviews (Camera) taken from a larger set of Epinions reviews also used by Bloom et al. [5], for a total of 4,700 texts split equally between positive and negative. Table 1 shows the performance of the English calculator with all features, and disabling the three types of valence shifters (negation, intensification and irrealis) and the extra weight on negative words. An asterisk (*) indicates that a chi-square test yielded significance at the $p<0.05$ level, as compared to the result with all features enabled. Whereas not all the differences are statistically significant, it does seem that the set of features that we have chosen has a positive effect on performance.

Table 1. Effects of disabling various features

	Percent Correct by Corpus			
Features	**Epinions**	**Movie**	**Camera**	**Total**
All	80.3	76.4	80.3	78.7
No Neg	75.8*	74.6	76.1*	75.4*
No Int	79.0	74.7	77.5*	76.5*
No Irreal	78.8	74.8	79.6	77.6
No Neg W	71.8*	75.6	71.5*	73.2*

3. The Spanish SO Calculator

Compared to English, Spanish is a highly inflected language, with gender and plural markers on nouns, as well as a rich system of verbal inflection (45 possible verb forms). In the English version of SO-CAL, the only external software we made use of was the Brill tagger [7]; lemmatization of noun and verbs was simple enough to be carried out during the calculation. For Spanish, we used a high-accuracy statistical tagger, the SVMTool [9], and we adapted a 500,000+ word lemma dictionary included in the FreeLing software package[1], which we used to both lemmatize the words and to add more detail to the basic verb tags assigned by SVMTool (each verb is lemmatized, but tagged with information about its tense and mood). We found that some sentiment-relevant words were not being lemmatized properly, so we also implemented a second layer of lemmatization within the calculator.

Most of the Python code written for the English version of SO-CAL could be reused. With regards to detecting negation, intensification, and modifier blocking, it was necessary to take into account the fact that in Spanish adjectives appear both before and (more commonly) after the noun. The most interesting difference was the fact that verb forms in Spanish provide irrealis information. In particular, the conditional tense and the imperative and subjunctive moods often serve to indicate that the situation being referred to is not in fact the case. Thus, in Spanish we used a mixture of word and inflection-based irrealis blocking, using the same words as the English version whenever possible.

We built new Spanish dictionaries, including dictionaries for adjectives, nouns, verbs, adverbs and intensifiers. For intensifiers, given the fact that they are closed-class and highly idiosyncratic, we simply created a new list of 157 expressions, based on the English list. For the open-class dictionaries, we tested three different methods of dictionary-building; we compare their performance on the Spanish corpus in Section 5.

The first set of dictionaries started with the English dictionaries for each part of speech, which we translated automatically into Spanish, preserving the semantic orientation value for each word. For the automatic translation we used, in turn, two different methods. The first was an online bilingual dictionary, from the site www.spanishdict.com. We extracted the first definition under the appropriate syntactic category, ignoring any cases where either the English or the Spanish were multi-word expressions. The second automatic translation method involved simply plugging our English dictionaries into the Google translator and parsing the results.

For the second method of dictionary creation, we took the lists from Spanishdict.com and manually fixed entries that were obviously wrong. This involved mostly removing words in the wrong dictionary for their part of speech, but also changing some of the values (less than 10% for each dictionary). This hand-correction took a native speaker of Spanish about two hours to complete.

Finally, the third method consisted in creating all dictionaries from scratch. Our source corpora created for this project consists of reviews extracted from the Ciao.es review website. Following the basic format of the Epinions corpus, we collected 400 reviews from the domains of hotels, movies, music, phones, washing machines, books, cars, and computers. Each category

[1] http://garraf.epsevg.upc.es/freeling/

contained 50 reviews: 25 positive and 25 negative. Whenever possible, exactly two reviews, one positive and one negative, were taken for any particular product, so that the machine learning classifier described in Section 4.2 could not use names as sentiment clues.

We tagged the Spanish corpus collected from Ciao.es, and extracted all adjectives, nouns, adverbs and verbs. This resulted in large lists for each category (e.g., over 10,000 nouns). We manually pruned the lists, removing words that did not convey sentiment, misspelled and inflected words, and words with the wrong part of speech tag. Finally, semantic orientation values were assigned for each. This process took a native speaker of Spanish about 12 hours. We decided against a committee review of the Spanish dictionaries for the time being.

Another type of dictionary tested was a merging of the dictionaries created using the second and third methods, i.e., the automatically-created (but hand-fixed) dictionaries and the ones created from scratch (Ciao manual). We created two versions of these dictionaries, depending on whether we used the value from the Fixed Spanishdict.com or Ciao dictionary.

The dictionaries range from smallest (Spanishdict.com) to largest (Ciao+Fixed). The first one contains 1,160 adjectives, 979 nouns, 500 verbs and 422 adverbs. The combined dictionary has 2,049 adjectives, 1,324 nouns, 739 verbs, and 548 adverbs.

We performed a comparison of fully automated and fully manual methods, comparing the unedited Spanishdict.com dictionaries and the ones created by hand. We calculated the percentage of words in common, as a percentage of the size for the larger of the two sets (the Spanishdict.com dictionaries). The commonalities ranged from roughly 20% of the words for nouns to 41% for adjectives (i.e., 41%, or 480 of the hand-ranked adjectives were also found in the automatic dictionary). We also compared the values assigned to each word: The variance of the error ranged from 1.001 (verbs) to 1.518 (adjectives). Automatically translated dictionaries tend to include more formal words, whereas the ones created by hand include many more informal and slang words

4. Alternative approaches

4.1 Corpus translation

For translation, we used Google's web-based translation system. Google Translate (translate.google.com) uses phrase-based statistical machine translation. We used only one translator, but Bautin et al. [4] discuss the use of different Spanish translating systems, and Wan [18] compare Chinese machine translators; the latter found that Google gave the best performance, which is consistent with our preliminary testing of other systems.

4.2 Machine Learning

A popular approach to sentiment analysis has been the automatic training of a text classifier. Cross-linguistic sentiment detection seems particularly amenable to machine learning, since classifiers can be easily trained in any language. Following Pang et al. [13], we used Support Vector Machine (SVM) classifiers, built with the sequential minimal optimization algorithm included in the WEKA software suite [20], with a linear kernel and testing done with 10-fold cross-validation. We trained using unigram features that appeared at least four times in the dataset (the same cut-off was used by Pang and Lee [12]). To test the efficacy of the WEKA classifiers, we first trained a classifier on the full 2,000 text Polarity Dataset, a collection of balanced positive and negative movie reviews [12], comparing the cross-validated results with the baseline for SVM unigram classifiers on this dataset (before other improvements) given in Pang and Lee [12]. The difference (about 1%) was not statistically significant. It is worth noting that more recent work in SVM-based sentiment analysis has shown significant improvement on this baseline [19], however relevant resources are not available for Spanish.

In order to compare the classifier across languages, we trained separately on each of our two 400-text development corpora. In each case we used the output after pre-processing, with lemmatizing in the case of Spanish. In addition to basic unigrams we also tested unigrams with full POS tags and, for Spanish, partial tags (retaining word class but disregarding inflection such as number and person). The results were identical or in some cases worse than a simple unigram model.

5. Evaluation

We built two additional 400 text corpora, in English and Spanish, with the same basic constituency as the Epinions and Ciao Corpus discussed earlier. The English corpus (Epinions 2) is also from the Epinions site, while the Spanish corpus came from Dooyoo.es. This second set of texts for each language has never been used for training or development of any of our resources

All four corpora were translated using the appropriate Google translator, and for each version the accuracy identifying the polarity of reviews for all possible dictionaries and methods was tested. Note that when the corpus and the dictionary are the same language, the original version of the corpus is used, and when the corpus and the dictionary are in different languages, we use the translated version. The results are given in Table 2.

There are a number of clear patterns in Table 2. First, for the original Spanish versions, the translated Spanish dictionaries, taken together, do poorly compared to the versions of the dictionaries derived from actual Spanish texts; this is significant at the $p<0.05$ level for all possible dictionary combinations (all significance results are derived from chi-square tests). For Spanish, including words from translated dictionaries has little or no benefit. The opposite is true for Spanish translations of English texts, where the Ciao (manual) dictionary performance is low, and performance improves dramatically with the addition of translated (although manually fixed) resources; in the case of the Epinions 2 corpus, this improvement is significant ($p<0.05$). We attribute this to the fact that translated texts and translated dictionaries "speak the same language"; translated English corpora

Table 2. Accuracy of polarity detection for various corpora and methods

Method		Corpus				
		English		Spanish		Overall
SO Calculator	**Dictionary**	**Epinions**	**Epinions2**	**Ciao**	**Dooyoo**	
English	English SO-CAL	**80.25**	**79.75**	72.50	**73.50**	**76.50**
Spanish	Google-translated	66.00	68.50	66.75	66.50	66.50
Spanish	Spanishdict.com	68.75	68.00	67.25	67.25	67.94
Spanish	Fixed Spanishdict.com	69.25	69.75	68.25	68.00	68.81
Spanish	Ciao (manual)	66.00	67.50	**74.50**	72.00	70.00
Spanish	Ciao + Fixed Combined, Ciao value preferred	68.75	**72.50**	74.25	72.25	**71.93**
Spanish	Ciao + Fixed Combined, Fixed value preferred	69.50	68.75	73.50	70.75	70.87
Support Vector Machine, English versions		76.50	71.50	72.00	64.75	71.25
Support Vector Machine, Spanish versions		71.50	68.75	72.25	69.75	70.56

are unlikely to contain the colloquial Spanish found in the Ciao dictionary, and are more likely to contain the kind of formal language we saw in our translated dictionaries.

Turning now to machine learning methods, the SVM classifiers show the worse performance overall, however only the difference seen in the Epinions 2 corpus is significant (at the $p<0.01$ level). The relatively poor performance of the SVM classifier in this case can be attributed to the small size of the training set and the heterogeneity of the corpora; SVM classifiers have been shown to have poor cross-domain performance in text sentiment tasks [2], a problem that can be remedied somewhat by integrating a lexicon-based system [1].

The numbers in Table 2 do not indicate a clear winner with respect to the performance of Spanish SO-CAL as compared to English SO-CAL with translated texts, although it is clear that translating English texts into Spanish is, at present, a bad approach ($p<0.01$). The totals for all corpora for each method suggest that Spanish SO-CAL is performing well below English SO-CAL ($p<0.01$).

Table 3 summarizes the effects of translation. Original refers to all the 1,600 original versions and Translated to all 1,600 translated versions. For SO calculation, we use the best performing dictionary in the relevant language.

Table 3. Accuracy for translated/original corpora

Method	Texts	Accuracy
SO Calculation	Original	76.62
	Translated	71.81
SVM	Original	72.56
	Translated	69.25

Table 3 shows a general deficit for translated texts; for SO calculation, this is significant at the $p<0.01$ level. The fact that it is also visible in SVMs (which are not subject to dictionary biases) suggests that it is a general phenomenon. One potential criticism here is our use of corpora whose words were the basis for our dictionary, unfairly providing two of the four original corpora with high coverage which would not pass to the translations. Indeed, there is some evidence in Table 3 to suggest that these high coverage corpora do outperform their low coverage counterparts to some degree in relevant dictionaries (compared with the Subjective dictionary, for instance); in general, though, there were no significant differences among same-language corpora tested using the same dictionary. Note also that using high-coverage corpora is not analogous to testing and training on the same corpora, since words are rated for SO independently of the texts in which they appear.

6. Related Work

Wan [18] created a hybrid classifier which combined the scores from a Chinese lexicon-based system and an English lexicon-based system (with translated texts). In contrast to our results, his Chinese lexicon-based system performed quite poorly compared to the English system. Similar to our results, Chinese lexicons created by translating English lexicons did not help performance.

Although they are concerned with sentence level subjectivity instead of text-level polarity, the work of Mihalcea et al. [11] is quite relevant, since their focus, like ours, is on exploring ways to deriving new resources from existing resources for English. In adapting subjectivity cues to Romanian, they also saw limited benefits to straight translation of dictionaries, but obtained promising results from the projection of English annotations into Romanian.

Bautin et al. [4] used online resources from multiple languages, including Spanish, into English, using the output from an existing sentiment analyzer to track attitudes in different language communities. Yao et al. [21] made use of a bilingual lexicon to build a Chinese sentiment dictionary using English glosses. Lexicon-based sentiment analysis has also been pursued independently in a number of East Asian languages, including Japanese [16], Chinese [22], and Korean [8]. As far as we know, ours is the first Spanish SO calculator. Banea et al. [3] report on work in Spanish, but theirs is a subjectivity classification task.

In terms of approaches to calculation of text level sentiment in English, the work of Kennedy and Inkpen [10] is the most directly comparable. Their main focus was the comparison of lexicon-based versus machine

learning approaches; in contrast to our results, they found that performance of their semantic model was significantly below that of an SVM classifier.

To facilitate comparisons with other approaches, the corpora and some of the resources described in the paper are available[2].

7. Conclusion

The surge in attention paid to automated analysis of text sentiment has largely been focused on English. In this paper, we have discussed how to adapt an existing English semantic orientation system to Spanish while at the same time comparing several alternative approaches.

Our results indicate that SVMs, at least the fairly simple SVMs we have tested here, do not do very well in our Spanish corpora. There are a number of obvious reasons for this, and our rejection of SVMs is far from decisive; on the contrary, machine learning might be useful, for instance, in identifying parts of the text that should be disregarded during the SO calculation [12].

For calculation of semantic orientation using lexicons, translation of any kind seems to come with a price, even between closely related languages such as English and Spanish. Our Spanish SO calculator (SO-CAL) is clearly inferior to our English SO-CAL, probably the result of a number of factors, including a small, preliminary dictionary, and a need for additional adaptation to a new language. Translating our English dictionary also seems to result in significant semantic loss, at least for original Spanish texts. Although performance of Spanish texts translated into English is comparable to native SO-CAL performance, the overall accuracy of translated texts in both English and Spanish suggests that there is 3-5% performance cost for any (automated) translation. This, together with the fact that translation seems to have a disruptive effect on previous reliable improvements, as well as the relatively small time investment required to develop Spanish SO-CAL, lead us to conclude that there is value in pursuing the development of language-specific resources, notwithstanding new breakthroughs in machine translation.

8. Acknowledgments

This work was supported by a NSERC Discovery Grant (261104-2008) to Maite Taboada.

9. References

[1] A. Andreevskaia and S. Bergler. When specialists and generalists work together: Domain dependence in sentiment tagging. Proc. of 46th ACL. Columbus, OH, 2008.

[2] A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: A case study. Proc. of RANLP. Borovets, Bulgaria, 2005.

[3] C. Banea, R. Mihalcea, J. Wiebe and S. Hassan. Multilingual subjectivity analysis using machine translation. Proc. of EMNLP. Honolulu, 2008.

[4] M. Bautin, L. Vijayarenu and S. Skiena. International sentiment analysis for news and blogs. Proc. of 3rd AAAI International Conference on Weblogs and Social Media. San Jose, CA, 2008.

[5] K. Bloom, G. Navendu and S. Argamon. Extracting appraisal expressions. Proc. of HLT/NAACL. Rochester, NY, 2007.

[6] J.D. Boucher and C.E. Osgood. The Pollyanna hypothesis. Journal of Verbal Learning and Verbal Behaviour 8: 1-8, 1969.

[7] E. Brill. A simple rule-based part of speech tagger. Proc. of 3rd Conference on Applied Natural Language Processing. Trento, Italy, 1992.

[8] Y.H. Cho and K.J. Lee. Automatic affect recognition using natural language processing techniques and manually built affect lexicon. IEICE Transactions on Information and Systems 89(12): 2964-2971, 2006.

[9] J. Giménez and L. Màrquez. SVMTool: A general POS tagger generator based on support vector machines. Proc. of Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal, 2004.

[10] A. Kennedy and D. Inkpen. Sentiment classification of movie and product reviews using contextual valence shifters. Computational Intelligence 22(2): 110-125, 2006.

[11] R. Mihalcea, C. Banea and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. Proc. of ACL. Prague, Czech Republic, 2007.

[12] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proc. of ACL. Barcelona, Spain, 2004.

[13] B. Pang, L. Lee and S. Vaithyanathan. Thumbs up? Sentiment classification using Machine Learning techniques. Proc. of EMNLP, 2002.

[14] L. Polanyi and A. Zaenen. Contextual valence shifters. In Computing Attitude and Affect in Text: Theory and Applications, J.G. Shanahan, Y. Qu, and J. Wiebe, Eds. Springer: Dordrecht, pp. 1-10, 2006.

[15] P.J. Stone. Thematic text analysis: New agendas for analyzing text content. In Text Analysis for the Social Sciences, C. Roberts, Ed. Lawrence Erlbaum: Mahwah, NJ, 1997.

[16] H. Takamura, T. Inui and M. Okumura. Extracting semantic orientations of words using spin model. Proc. of ACL. Ann Arbor, 2005.

[17] P. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proc. of ACL, 2002.

[18] X. Wan. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. Proc of EMNLP. Honolulu, 2008.

[19] C. Whitelaw, N. Garg and S. Argamon. Using Appraisal groups for sentiment analysis. Proc. of ACM SIGIR Conference on Information and Knowledge Management (CIKM 2005). Bremen, Germany, 2005.

[20] I.H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[21] J. Yao, G. Wu, J. Liu and Y. Zheng. Using bilingual lexicon to judge sentiment orientation of Chinese words. Proc. of 6th International Conference on Computer and Information Technology (CIT'06). Seoul, Korea, 2006.

[22] Q. Ye, B. Lin and Y.-J. Li. Sentiment classification for Chinese reviews: A comparison between SVM and semantic approaches. Fourth Int. Conference on Machine Learning and Cybernetics. Guangzhou, China, 2005.

[2] http://www.sfu.ca/~mtaboada/research/nserc-project.html