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 (,

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

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 ol,