LING 575: Intermediate Project Report

Claire Jaja
University of Washington
Seattle, WA
cjaja@uw.edu

Andrea Kahn
University of Washington
Seattle, WA
amkahn@uw.edu

Abstract

In this paper, we propose an approach to compare the effectiveness of supposedly language-agnostic sentiment analysis techniques such as machine learning algorithms on text in different languages. In addition, we plan to compare results obtained when using resources such as sentiment lexicons that are pivoted or translated from English versus using resources that are developed in the test language.

1 Introduction

Sentiment analysis techniques are typically developed on English text. However, there is a proliferation of text in other languages as well, which could benefit from sentiment analysis. Current approaches to sentiment analysis in languages other than English often involve automatically translating the text into English, which is likely to introduce errors, as well as increasing runtimes. Furthermore, we suspect that syntactic variation between languages means that different features may be more useful for different languages.

Some techniques build on machine learning algorithms that can easily be applied to other languages. Others rely on manually developed resources such as sentiment lexicons, which are language-specific; often lexicons in languages other than English are translated or pivoted from English rather than developed for that language.

In this paper, we propose an approach to compare the effectiveness of supposedly language-agnostic techniques such as machine learning algorithms on text in different languages. In addition, we plan to compare results obtained when using resources such as sentiment lexicons that are pivoted or translated from English versus using resources that are developed in the test language.

2 Related Work

Balahur Dobrescu (2011) compares the results of performing sentiment analysis and opinion mining on different text types using a variety of multilingual resources, many of them hand-built. She allegedly obtains ”good” results even on texts in languages for which no resources are available and either the texts or resources need to be translated, though we plan to take a closer look.

Brooke et al. (2009) adapt various English resources, including semantic orientation calculators and dictionaries, for Spanish sentiment analysis. They also compare the results of using machine translation to translate the Spanish text into English, as well as using a language-independent machine learning algorithm (SVM classification) to classify the Spanish text. They conclude that the results obtained using adapted resources are not comparable to those using language-specific resources.

Vilares et al. (2013b) describes a system for polarity classification of Spanish texts that uses dependency parsing to access the syntactic structure of the sentences to be classified. This paper is unique in that it diverges from the shallow approaches taken in most polarity classification papers, and whether or not we are able to employ deep processing techniques, we hope to learn more about challenges that may be specific to Spanish sentiment analysis from the approach and results presented here.

Other papers describing related work that we hope to study and draw from include Boiy and Moens (2009), Cruz Mata (2011), Fernández Anta (2012), Pang and Lee (2004), Pang et al. (2002), Urizar et al. (2012), Vilares et al. (2013a), and Zhang et al. (2009).

3 Data

For our experiments, we are using comparable corpora in multiple languages.

Firstly, we use the polarity dataset v2.0 described in Pang and Lee (2004), which consists of 1000 positive and 1000 negative pre-processed reviews from IMDb. For a comparable dataset in another language, we use CorpusCine Reviews (Cruz Mata, 2011), a collection of 3,878 movie reviews written in Spanish from the muchocine.net web page. Each review has a rating between one (most negative) and five (most positive) stars. For our purposes, we discard the three star reviews and classify the one and two star reviews as negative and the four and five star reviews as positive. This leaves us with 1,274 negative reviews and 1,351 positive reviews.

Additionally, we have a corpus of 1,590 English quotations from newspaper articles annotated for polarity and a comparable corpus of 2,387 German quotations from newspaper articles; both of these were annotated based on the same annotation criteria (Balahur Dobrescu, 2011).

4 Approach

We employ the Mallet toolkit to train and test polarity classifiers on these two datasets. We plan to test a variety of classification algorithms and features, in order to determine if certain ones provide better results for one language or another. Additionally, we plan to do a linguistic error analysis, to tease apart the impact of the language differences (in this vein, Andrea hopes to use this course as a linguistics elective, while Claire is using it as a computational linguistics elective).

5 Results

We have implemented a baseline system on both movie review datasets using a MaxEnt classifier and unigram bag of word features. For each dataset, we use 90% of the reviews for training and 10% for testing; this gives us a training set of 900 positive and 900 negative reviews for the IMDb corpus and 1,215 positive and 1,146 negative reviews for the CorpusCine corpus. There are 100 positive and 100 negative reviews in the IMDb test set and 136 positive and 128 negative reviews in the CorpusCine test set. The results for this system are presented below.

Classifier	**features**	**IMDb**	**CorpusCine**
MaxEnt	unigram	88.00%	83.71%

Table 1: Baseline results. Test accuracy for IMDb and CorpusCine.

label	*negative*	*positive*
negative	86	14
positive	10	90

Table 2: IMDb confusion matrix. Row = true, column = predicted.

label	*negative*	*positive*
negative	103	25
positive	18	118

Table 3: CorpusCine confusion matrix. Row = true, column = predicted.

6 Discussion

The task we approach in this project has many challenges. In addition to the challenge of polarity classification, finding appropriately parallel corpora in two languages and performing preprocessing on the data so that the same techniques could be applied to data from both corpora proved difficult. Likewise, finding parallel sentiment lexicons that are both pivoted/translated and native (in multiple languages) will be a challenge. Finally, we will be faced with the task of analyzing results to determine whether differing results in different languages and/or using different resources are due to differing syntactic properties of the languages.

Our next steps include cutting back the CorpusCine train and test sets to have the same number of reviews as the IMDb train and test sets to make the results as directly comparable as possible. To this end, we will also compare review lengths (i.e. average, minimum, and maximum number of tokens and types). After that, we hope to implement n-fold cross validation of our results (most likely with n=10).

Once our datasets are fully established, we will test different classification algorithms and features, as well as a variety of external resources (e.g. polarity lexicons, both native and pivoted/translated).

7 Conclusion

References

Alexandra Balahur Dobrescu. 2011. *Methods and Resources for Sentiment Analysis in Multilingual Doc-*

uments of Different Text Types. Ph.D. thesis, Universidad de Alicante.

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558.

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From English to Spanish. In *International Conference RANLP*, pages 50–54.

Fermin Luis Cruz Mata. 2011. *Extraccíon de Opiniones sobre Caracterısticas: Un Enfoque Práctico Adaptable al Dominio*. Ph.D. thesis, Universidad de Sevilla.

Antonio Fernández Anta. 2012. Techniques for sentiment analysis and topic detection of Spanish tweets: preliminary report.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Xabier Saralegi Urizar, Elhuyar Fundazioa, and Inaki San Vicente Roncal. 2012. Tass: Detecting sentiments in Spanish tweets. *TASS 2012 Working Notes*.

David Vilares, Miguel Ángel Alonso, and Carlos Gómez-Rodríguez. 2013a. Supervised polarity classification of Spanish tweets based on linguistic knowledge. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 169–172. ACM.

David Vilares, Miguel Ángel Alonso, and Carlos Gómez-Rodríguez. 2013b. A syntactic approach for opinion mining on Spanish reviews. *Natural Language Engineering*, pages 1–25.

Changli Zhang, Daniel Zeng, Jiexun Li, Fei-Yue Wang, and Wanli Zuo. 2009. Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12):2474–2487.