Sense-level Subjectivity in a Multilingual Setting

Carmen Banea
University of North Texas
carmenbanea@my.unt.edu

Rada Mihalcea
University of North Texas
rada@cs.unt.edu

Janyce Wiebe
University of Pittsburgh
wiebe@cs.pitt.edu

Abstract

This paper explores the ability of senses aligned across languages to carry coherent subjectivity information. We start out with a manual annotation study, and then seek to create an automatic framework to determine subjectivity labeling for unseen senses. We identify two methods that are able to incorporate subjectivity information originating from different languages, namely co-training and multilingual vector spaces, and show that for this task the latter method is better suited and obtains superior results.

1 Introduction

Following the terminology proposed by (Wiebe et al., 2005), subjectivity and sentiment analysis focuses on the automatic identification of private states (opinions, emotions, sentiments, etc.) in natural language. While subjectivity classification labels text as either subjective or objective, sentiment or polarity classification further classifies subjective text as either positive, negative or neutral.

To date, a large number of text processing applications have used techniques for automatic sentiment and subjectivity analysis, including automatic expressive text-to-speech synthesis (Alm et al., 1990), tracking sentiment timelines in on-line forums and news (Balog et al., 2006; Lloyd et al., 2005), and mining opinions from product reviews (Hu and Liu, 2004). In many natural language processing tasks, subjectivity and sentiment classification has been used as a first phase filtering to generate more viable data. Research that benefited from this additional layering ranges from question answering (Yu and Hatzivassiloglou, 2003), to conversation summarization (Carenini et al., 2008), text semantic analysis (Wiebe and Mihal-

cea, 2006; Esuli and Sebastiani, 2006a) and lexical substitution (Su and Markert, 2010).

While research in English has underlined that the most robust subjectivity delineation occurs at sense and not at word level (Wiebe and Mihalcea, 2006), we are not aware of this consideration impacting research in other languages. For this reason, in this work we seek to analyze how subjectivity is maintained across sense aligned resources, and identify ways in which subjectivity at sense level may be employed in a multilingual framework to provide a strengthened automatic sense-level classification.

2 Related Work

Recently, resources and tools for sentiment analysis developed for English have been used as a starting point to build resources in other languages, via cross-lingual projections or monolingual and multilingual bootstrapping. Several directions were followed, focused on leveraging annotation schemes, lexicons, corpora and automated annotation systems. English annotation schemes developed for opinionated text lays the groundwork for research carried out by (Esuli et al., 2008) when annotating expressions of private state in Italian or by (Maks and Vossen, 2010) in Dutch. Sentiment and subjectivity lexicons such as the one included with the OpinionFinder distribution (Wiebe and Riloff, 2005), the General Inquirer (Stone et al., 1967), or the SentiWordNet (Esuli and Sebastiani, 2006b) were transferred into Chinese (Ku et al., 2006; Wu, 2008) and into Romanian (Mihalcea et al., 2007). English corpora manually annotated for subjectivity or sentiment such as MPQA (Wiebe et al., 2005), or the multi-domain sentiment classification corpus (Blitzer et al., 2007) were subjected to experiments in Spanish, Romanian, or Chinese upon automatic translation by (Banea et al., 2008b; Wan, 2009). Furthermore, tools developed for English were used to determine sentiment

or subjectivity labeling for a given target language by transferring the text to English and applying an English classifier on the resulting data. The labels were then transferred back into the target language (Bautin et al., 2008; Banea et al., 2008b). These experiments are carried out in Arabic, Chinese, French, German, Japanese, Spanish, Romanian.

We are not aware of research that has considered leveraging subjectivity at word sense level, yet, in terms of methodology, the work closest to ours is the one proposed by (Wan, 2009), who constructs a polarity co-training system by using the multilingual views obtained through the automatic translation of product-reviews into Chinese and English. Unlike (Wan, 2009), we do not use any machine translation, and the labels employed are directly assigned by the annotators and not inferred based on stars. (Banea et al., 2008a) present a method to learn sentence level subjectivity by training classifiers on multilingual feature spaces and show that when considering features from multiple languages, the classification accuracy improves, even above that of the source language. We expand this method to allow for bootstrapping, thus enabling additional samples to be classified.

3 Sense Level Subjectivity Consistency Across Languages

While most multilingual research to date has focused on word, fragment, or document level subjectivity, this work seeks to examine sense-level subjectivity across languages. We aim to answer two questions. First, if we have a resource such as WordNet (Miller, 1995) aligned at sense level in two languages, is the subjectivity content consistent across equivalent senses in the two languages? Second, can we use a multilingual learning mechanism to automatically predict the subjectivity label of senses? We examine the first question in Section 3.1, and propose a framework for multilingual learning that responds to the second question in Section 3.2.

3.1 Annotation Study

For the purpose of this study we consider the English (Miller, 1995) and the Romanian (Tufiş et al., 2006) versions of WordNet, which contain

117659[1] and 58725[2] synsets, respectively. Both lexical resources are aligned at *synset* level, which represents a basic unit of meaning.

In order to add subjectivity information to this structure, we use the English annotated data from (Wiebe and Mihalcea, 2006) and (Akkaya et al., 2009), as well as a list of 48 additional words, for a total of 134 words encompassing 630 senses manually annotated for subjectivity. This data was then annotated by a native speaker of Romanian (who participated in previous subjectivity annotations studies) who was only presented with the gloss and the synset of each given sense from the Romanian WordNet. The agreement with the English annotations ranged from 90% (for the (Wiebe and Mihalcea, 2006) dataset) to 84% (for the (Akkaya et al., 2009) dataset), implying that subjectivity can strongly transfer across senses given manually aligned resources in different languages. However, we encountered several situations that may interfere with the subjective content of a sense, which are further explained below.

3.1.1 Differences between Languages

There were several examples where the subjectivity label changed between languages. Let us consider the following definitions of the fourth sense of the noun *argument* listed in Table 1. While this sense of *argument* is marked in the English data as subjective, the Romanian gloss and synset denote a "direct summary," which by definition disallows the expression of any subjective perspective. Therefore, in Romanian this sense is objective.

A similar scenario is posed by the fourth sense of the verb *decide* (see Table 1). While the English sense is labeled as objective, the Romanian sense directly implies a subjective decision, and therefore acquires a subjective label.

3.1.2 WordNet Granularity

In several cases, the same sense in WordNet may have both subjective and objective meanings. To exemplify, let us consider the first sense of the adjective *free*:

En gloss: not limited or hampered; not under compulsion or restraint; "free enterprise"; "a

---

	English	Romanian
argument		
Gloss	a summary of the subject or plot of a literary work or play or movie "the editor added the argument to the poem"	redare-prezentare pe scurt- scrisă sau orală- a ideilor unei lucrări- ale unei expuneri etc. (*translation*) short summary, oral or in writing, of the ideas presented in a literary work
Synset	argument, literary argument	rezumat (*translation*) summary
decide		
Gloss	influence or determine "The vote in New Hampshire often decides the outcome of the Presidential election"	a exercita o influenţă - a determina (*translation*) to exercise influence - to determine
Synset	decide	influenţa; decide; hotărî (*translation*) influence; decide; determine

Table 1: *Differences between languages.* Definitions and synonyms of the fourth sense of the noun *argument* and the fourth sense of verb *decide* as provided by the English and Romanian WordNets; for Romanian we also provide the manual translation into English.

free port"; "a free country"; "I have an hour free"; "free will"; "free of racism"; "feel free to stay as long as you wish"; "a free choice"
Ro gloss: (Despre oameni) Care are posibilitatea de a acţiona după voinţa sa - de a face sau de a nu face ceva; (*translation*) (About people) Someone who can act according to his will - who can do or not do something

While the English sense can have both subjective and objective uses, the Romanian sense is subjective, as it further enforces the constraint that the context of the word should refer to people.

From these examples, we notice that a perfect sense to sense mapping among languages is impossible, as a particular sense may denote additional meanings and uses in one language compared to another, thus rendering a perfect parallel sense boundary permeable. However, for about 90% of the senses the subjective meaning does hold across languages, implying that this information could be leveraged in an automatic fashion to provide additional clues for the subjectivity labelling of unseen senses.

3.2 Multilingual Subjectivity Sense Learning

In this section we explore ways to use a multilingual learning mechanism to automatically predict the subjectivity of a word sense. We are experimenting with two different methods, one based on

co-training using monolingual feature spaces, and one based on machine learning applied to a multilingual vector space.

We start by considering the intersection of the Romanian and English WordNets, so that we can have equivalent definitions in both languages. We then generate vector representations for two monolingual models (one in English and one in Romanian), and one multilingual model (comprising both Romanian and English features). These are composed of unigrams extracted from the synset and the gloss of a given sense, appended with a binary weight. The synset is stripped of any sense identifying features in order not to favor the classifier. To exemplify, we provide below the sparse vector representation of the fourth sense of the noun *argument* (see Table 1):

English vector: $<a_{en}$ 1, summary 1, of 1, the 1, subject 1, or 1, plot 1, literary 1, work 1, play 1, movie 1, editor 1, added 1, argument 1, to 1, poem 1>
Romanian vector: <redare 1, prezentare 1, pe 1, scurt 1, scrisa 1, orala 1, a_{ro} 1, ideilor 1, unei 1, lucrari 1, ale 1, expuneri 1, etc 1, rezumat 1>
Multilingual vector: $<a_{en}$ 1, summary 1, of 1, the 1, subject 1, or 1, plot 1, literary 1, work 1, play 1, movie 1, editor 1, added 1, argument 1, to 1, poem 1, redare 1, prezentare 1, pe 1, scurt 1, scrisa 1, orala 1, a_{ro} 1, ideilor 1, unei 1, lucrari 1, ale 1, expuneri 1, etc 1, rezumat 1>

In the first method, based on the co-training algorithm proposed by (Wan, 2009), we consider the manually annotated training data in each of the languages individually, and we learn two monolingual classifiers (see Figure 1). We then allow the machine learners to individually predict a class for every sample in the unlabeled data, and at every iteration create a set with the top n most confident examples where both classifiers agree, and their confidence is higher than a given threshold. As long as the set has at least one sample, at the next iteration the monolingual English vectors and the aligned Romanian vectors are added to their respective training set with the newly predicted label, and removed from the test data. The process repeats until no confident examples can be added. Although the method differs from the original co-training mechanism proposed by (Blum and Mitchell, 1998), since it enforces that the classifiers agree before adding their predictions to the next train set, we believe this was a necessary modification given the low accuracy attained by the Romanian classifier by itself (68%). Through this additional agreement constraint, we ensure that only samples that have a high probability of being labeled correctly are added, therefore reducing noise propagation across iterations. At the same time, we are able to learn new information from the features co-occurring with those that participated in the previous classification step.

For the second method, we create a multilingual feature space based on the model proposed in (Banea et al., 2010). Instead of using the monolingual vectors described above, we enrich the feature space by merging together two aligned vector space representations (see the multilingual vector example above), thus allowing the system to simultaneously use both Romanian and English features in order to decide the subjectivity of a given sense. At every iteration we select the most confident n samples, and add them to the training set, while discarding them from the test set for the next iteration.

For all the experiments presented in this paper we use support vector machines (the LibSVM implementation (Fan et al., 2005)) with default parameters and probability estimates enabled. As we are interested in an accurate classification of the senses, we chose a threshold level of 0.8, and at

every iteration we add the most confident $n = 40$ samples to the previous training set.
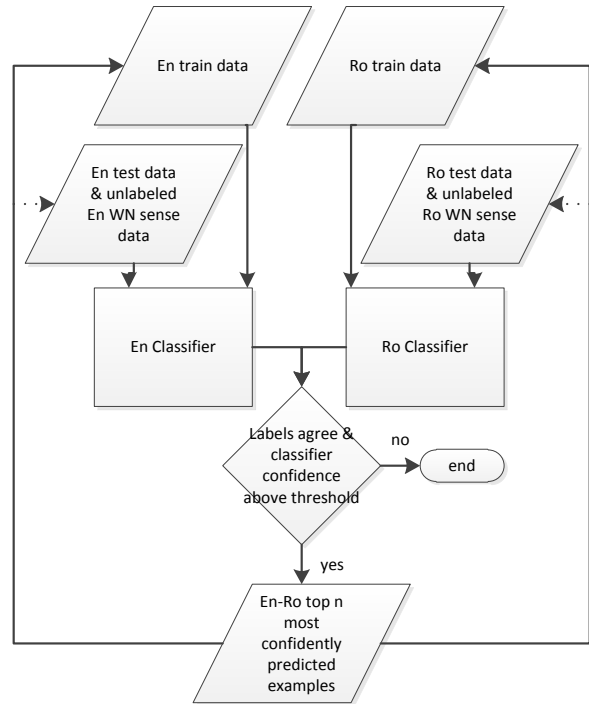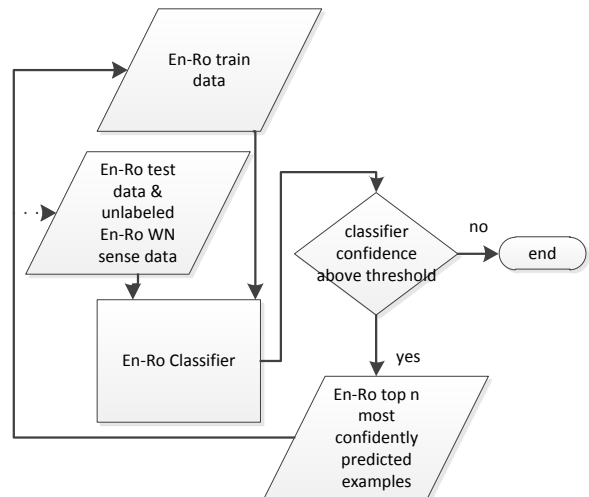

Figure 1: Co-training

Figure 2: Multilingual bootstrapping

3.2.1 Datasets

We use the manually annotated data described in Section 3.1, and we filter out 20 examples that were labeled as both objective and subjective, since they could confuse the classifiers and prevent them from making strong predictions. We then split the labeled data into three subsets to enable a three-fold cross validation. Note that we enforce that all the senses belonging to a given word

be found in either the test or the training set, but never in both. This was done to ensure that the classifier would not have an unfair advantage due to finding similar senses in the training data. For this reason, the fold sizes are not perfectly equal. Furthermore, for every fold, each iteration is evaluated on the immutable test set corresponding to that fold, which has manually assigned labels in English and Romanian. In order to generate a running test set, which is modified after every iteration, we append the remaining unlabeled WordNet senses to the corresponding test set for the fold (see Figures 1 and 2).

3.2.2 Results and Discussions

Figure 3 presents the results obtained using the monolingual co-training algorithm over 40 iterations. The accuracies obtained at position 0 represent the baseline for a simple monolingual classifier with no co-training. Unlike the increasing accuracy with the number of iterations obtained by (Wan, 2009) when applying a similar method to sentiment classification of reviews, we were unable to surpass these baselines. We attribute this behavior to the small size of the training set (approximately 400 samples in our case versus 8000 product reviews in (Wan, 2009)) and the type of data itself (product reviews are longer and often contain a full paragraph of text, while senses may comprise an average of ten words). The overall accuracy is slowly decreasing from 0.73 to 0.62 for English and from 0.68 to 0.54 for Romanian. The same trend is observed for class precision, recall and F-measure.

When employing a simple SVM classifier trained on a multilingual space, the accuracy increases from 0.73 for English and 0.62 for Romanian to 0.76 when both languages are simultaneously used, thus providing an error reduction of 11.34% and 25.74% with respect to the monolingual English and Romanian models, respectively. Since the English WordNet is more complete (longer glosses and richer synsets), its corresponding monolingual model is able to capture sufficient information and thus provide a robust subjectivity classification on its own. However, upon training on a multilingual representation of the data, features from both languages synergistically work together to achieve better results than what would be individually possible. These results further confirm the improving trend we noticed in (Banea et al., 2010) when training classifiers on

incrementally more languages.

We also attempted to bootstrap the multilingual classifier (see Figure 4), but its performance degrades faster than when using the co-training method, and after only 3 iterations the confidence of the classifier drops below the threshold and the process terminates. It may be beneficial to add fewer instances to the training set at each iteration in order to introduce less noise and thus obtain a more robust classifier. This is a setting that we intend to explore in the future, however for the current experiments, in order to equitably compare the two methods, we kept all the parameters equal.
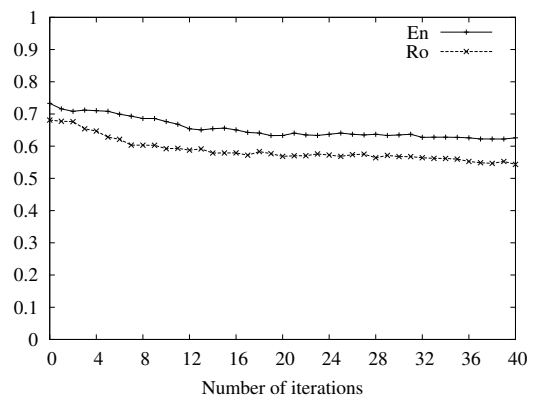

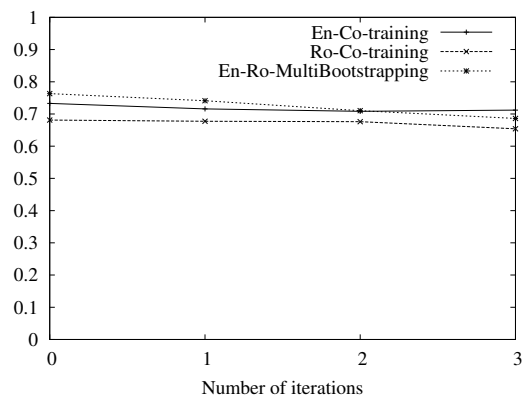Figure 3: Macro-accuracy for co-training

Figure 4: Macro-accuracy for multilingual bootstrapping versus monolingual co-training

4 Conclusion

We performed a manual annotation study for subjectivity at sense level and we showed that the subjectivity content of a sense does carry across language boundaries in about 90% of the cases, implying that this information is robust enough to be

learned automatically. We then proposed and applied a framework that is able to jointly exploit the subjectivity information originating from multiple languages. We demonstrated that a multilingual feature space is able to capture more information and outperform a monolingual based model, suggesting that future research should use a similar representation.

Acknowledgments

References

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity Word Sense Disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 190–199, Singapore, August.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 1990. Emotions from text: machine learning for text-based emotion prediction. *Intelligence*.

Krisztian Balog, Gilad Mishne, and Maarten De Rijke. 2006. Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.

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.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual Subjectivity: Are More Languages Better ? In *Proceedings of the International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2008)*, Seattle, Washington.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational (ACL-2007)*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, pages 92–100. Morgan Kaufmann.

Giuseppe Carenini, Raymond T Ng, and Xiaodong Zhou. 2008. Summarizing Emails with Conversational Cohesion and Subjectivity. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL- HLT 2008)*, pages 353–361, Columbus, Ohio.

Andrea Esuli and Fabrizio Sebastiani. 2006a. Determining Term Subjectivity and Term Orientation for Opinion Mining. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, volume 2, pages 193–200, Trento, Italy.

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

Andrea Esuli, Fabrizio Sebastiani, and Ilaria C Urciuoli. 2008. Annotating Expressions of Opinion and Emotion in the Italian Content Annotation Bank. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC-2008)*, Marrakech, Morocco.

Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. 2005. Working set selection using the second order information for training SVM. *Journal of Machine Learning Research*, 6:1889—-1918.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (ACM-SIGKDD-2004)*, pages 168–177, Seattle, Washington.

Lun-wei Ku, Yu-ting Liang, and Hsin-hsi Chen. 2006. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, number 2001, Boston, Massachusetts.

Levon Lloyd, Dimitrios Kechagias, and Steven Skiena. 2005. Lydia : A System for Large-Scale News Analysis (Extended Abstract) News Analysis with Lydia. In *Lecture Notes in Computer Science*, pages 161–166. Springer, Berlin / Heidelberg.

Isa Maks and Piek Vossen. 2010. Annotation scheme and gold standard for Dutch subjective adjectives. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, pages 1327–1334, Valletta, Malta.

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.

George A. Miller. 1995. WordNet: a Lexical database for English. *Communications of the Association for Computing Machinery*, 38(11):39—-41.

Philip J Stone, Marshall S Smith, Daniel M Ogilivie, and Dexter C Dumphy. 1967. *The General Inquirer: A Computer Approach to Content Analysis. /*. The MIT Press, 1st edition.

Fangzhong Su and Katja Markert. 2010. Word sense subjectivity for cross-lingual lexical substitution. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT'10)*, pages 357—-360, Los Angeles, CA, USA.

Dan Tufiş, Verginica Mititelu Barbu, Luigi Bozianu, and Cătălin Mihăilă. 2006. Romanian Wordnet: current state, new developments and applications. In *Proceedings of the 3rd Conference of the Global WordNet Association (GWC'06)*, pages 337–344, Seogwipo, Jeju Island, Republic of Korea.

Xiaojun Wan. 2009. Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, Singapore.

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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.

Yejun Wu. 2008. Classifying attitude by topic aspect for English and Chinese document collections.

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