Applying Sentiment-oriented Sentence Filtering to Multilingual Review Classification

Takashi Inui and **Mikio Yamamoto**
Graduate School of Systems and Information Engineering
University of Tsukuba
1-1-1 Tenoudai, Tsukuba, Ibaraki 305-8573, JAPAN
{inui,myama}@cs.tsukuba.ac.jp

Abstract

A method for multilingual review classification is described. In this classification task, machine translation techniques are used to remove language gaps in the dataset, but many translation errors occur as a side-effect. These errors cause a decrease in the review classification performance. To resolve this problem, we introduce a sentiment-oriented sentence filtering module to the process of multilingual review classification. Experimental results showed that the proposed method achieved 81.7% classification accuracy for the evaluation data.

1 Introduction

People can nowadays easily disseminate information including their personal subjective opinions on products and services on the Internet. The massive amounts of this type of information are beneficial for both product companies and users who are planning to purchase and use the products. The information is mainly presented in a textual form, so in the research field of natural language processing, many researchers have focused on developing techniques for *sentiment analysis* (or *opinion mining*) (Pang and Lee, 2008; Tang et al., 2009).

One fundamental technique in sentiment analysis (opinion mining) is to classify review texts. Unlike the conventional topic-based text classification task, classifiers for review classification must discriminate between *positive* and *negative* aspects of opinions in a review text. In the review classification task, supervised machine learning methods such as Naive Bayes and Support Vector Machines have been mostly applied (Pang et al., 2002; Mullen and Collier, 2004; Whitelaw et al., 2005). These supervised approaches have achieved good performance, but they have a crucial issue: they require a large amount of labeled data, which involves the high cost of manual annotation.

Approaches to reduce or avoid the cost of annotation have been proposed, such as semi-supervised and substitutional data approaches. Semi-supervised approaches (e.g., that by Aue and Gamon (2005)) provide a simple solution by combining labeled and unlabeled data. Substitutional data approaches provide substitutional labeled data, available at low costs, instead of pure labeled data. The tasks of domain adaption (Blitzer et al., 2007) and multilingual text classification (Banea et al., 2008; Wan, 2009; Banea et al., 2010) are special cases of substitutional approaches.

In this paper, we examine the effectiveness of applying a sentence filtering module to multilingual document classification, especially to multilingual review classification. In multilingual review classification, machine translation techniques are usually used to remove language gaps in the dataset. But, even if one can use the state-of-the-art machine translation techniques, many translation errors occur as a side-effect. These errors cause a decrease in the review classification performance. In this study, to resolve this problem, we introduce a sentiment-oriented sentence filtering module to the process of multilingual review classification. we focus on the quality rather than the quantity of the training data, and attempt to filter out some worthless sentences from the dataset.

The rest of this paper is organized as follows. First, we provide an overview of multilingual review classification in Section 2. In addition, an issue essentially related to the task of multilingual review classification is presented. In Section 3, we explain our sentiment-oriented sentence filtering method. In Section 4, we report on our experi-

Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 51–58,
Chiang Mai, Thailand, November 13, 2011.

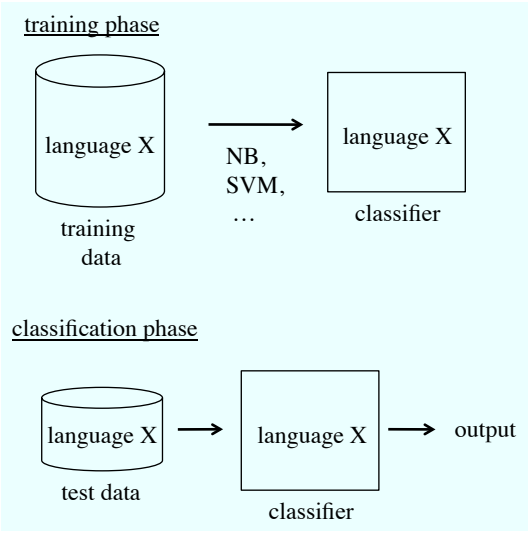

Figure 1: Monolingual review classification

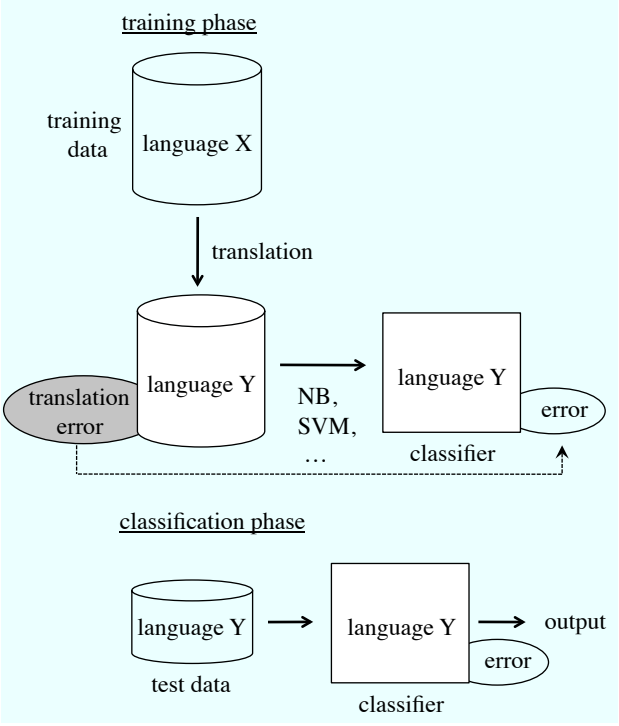

Figure 2: Multilingual review classification

ments investigating the effectiveness of applying our filtering method to multilingual review classification.

2 Multilingual Review Classification

2.1 Overview

Figure 1 shows an ordinary processing flow of text (review) classification with monolingual data. In a monolingual setting, in both the training phase and classification phase, text documents in the dataset are described in the same language (language X in Figure 1). Figure 2, in contrast, shows a multilingual setting for review classification. In this setting, text documents in the classification phase are described in a different language, Y, from X.

To remove the language gap between the training and test datasets, machine translation (MT) techniques are used in the training phase. By translating text documents in the dataset from the source language X into the target language Y, an MT system automatically generates a substitutional dataset in which text documents are described in the target language Y[1].

2.2 The issue

Here, the MT system succeeds in removing the language gap between X and Y. However, many translation errors occur in the dataset as a side-effect. In general, a text classifier uses information about word distributions over the dataset. When there are erroneous translations in the dataset, a situation is invoked in which the distributions of each word in the dataset differ between the training data and the test data. As a result, these errors cause increase of text classification errors indirectly (dotted line in Figure 2).

3 Applying Sentiment-oriented Sentence Filtering

In this section, our method for reducing the influences of translation errors is proposed. In the proposed method, documents translated by an MT system are then compressed by a sentiment-oriented sentence filtering module. We begin with discussion about our key idea of the proposed method, and then explain our sentiment-oriented sentence filtering.

3.1 Key idea

Consider the relationship between a labeled dataset for training a text classifier and its classification accuracy. In a general case, the larger the labeled training dataset, the better the performance of the text classifier. However, in the case of multilingual review classification, this relationship does not hold due to the translation errors be-

[1]Note that even though a small amount of original labeled documents is described (i.e., not translated) in language Y in general cases, this is omitted in Figure 2 for simplicity.

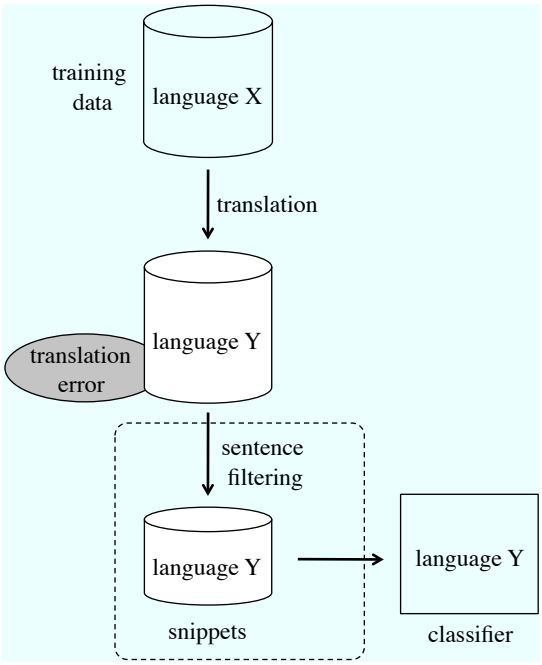

Figure 3: Multilingual review classification with sentence filtering

cause if the labeled data increase, the translation errors included in them may also increase. That is, the number of translation errors may be proportional to the number of labeled documents.

According to the above discussion, we focused our attention on the *quality* rather than the *quantity* of the labeled data. To achieve this change of focus, we introduce a sentence filtering module after the machine translation step. Figure 3 shows the training phase of multilingual review classification with the sentence filtering module. In the sentence filtering module, a translated document is compressed into a snippet consisting of important parts of the translated document for the review classification task. Since the generated snippet is shorter than the input document, and recalling that the number of translation errors may be proportional to the quantity of the dataset, applying sentence filtering should help to prevent errors being incorporated into the dataset.

3.2 Sentiment-oriented sentence filtering

Our sentence filtering module aims to generate text snippets by excluding translation errors from the input translated documents. To do so, we developed a sentiment-oriented sentence filtering method.

We need to develop criteria by which sentences should be extracted. The most direct approach is that *all sentences correctly translated are extracted and all remaining erroneous sentences are excluded.* This may work well, but it is infeasible because detecting whether a sentence is correctly translated is difficult.

Instead, we consider an alternative approach based on sentiment information. Pang et al. (2004) found that an important factor for a review classification task is whether each sentence in a document to be classified holds subjective aspects. Generally, subjective sentences contribute to the performance of review classification, while objective sentences do not. According to this finding, we adopted the following sentence filtering criteria: *all sentences holding subjective aspects are extracted and all remaining objective sentences are excluded.* We consider that objective sentences with translation errors are not only unnecessary but also harmful for the multilingual review classification.

In this study, we detect a sentence S_Y as holding subjective aspects when all the following conditions are fulfilled.

(1) S_Y includes at least one polarity word,

(2) A sentence S_X, which has a translation relation to S_Y, also includes at least one polarity word,

(3) All the polarity words in S_X and S_Y have the same sentiment polarity.

Condition (1) is commonly used in the field of sentiment analysis (Kim and Hovy, 2005). Conditions (2) and (3), on the other hand, are originally derived from the translation process in the multilingual review classification. By adding these two conditions, we achieve more robust subjectivity detection. Figure 4 shows an example of the sentence filtering process. Sentences S_{Y2} and S_{Y4} fulfill all the conditions and thus are extracted. Sentences S_{Y1} and S_{Y3} are excluded. S_{Y1} violates condition (1): it has no polarity words. S_{Y3} violates condition (3): although S_{Y3} has a negative polarity word, S_{X3} has a positive polarity word. In this example, one can see that the snippet generated keeps almost all the subjective information and also that it succeeds in eliminating parts of erroneous translations.

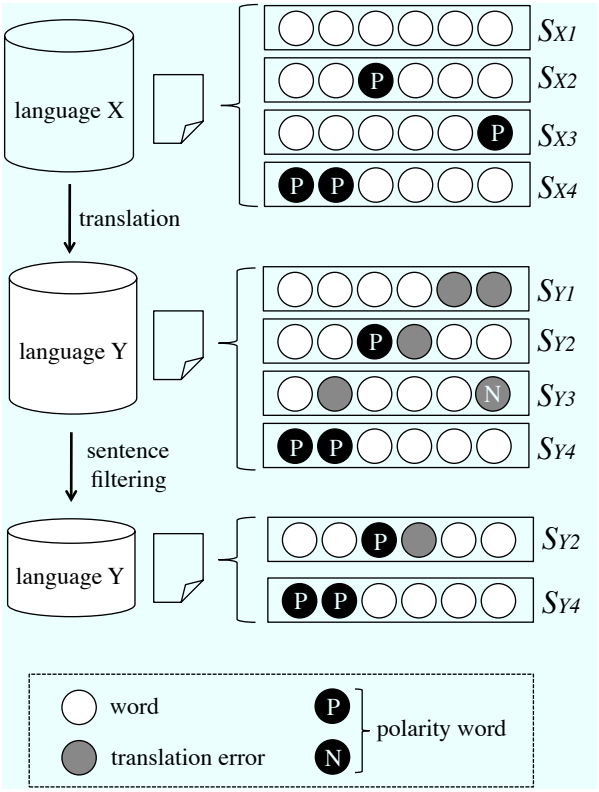

Figure 4: Example of sentence filtering process

4 Evaluation

We conducted experiments for investigating the effectiveness of applying our sentiment-oriented sentence filtering method to the multilingual review classification.

4.1 Experimental settings

4.1.1 Multilingual review classification methods

Review classification methods that enable handling of multilingual data have been proposed. We adopted those proposed by Banea et al. (2008) and Wan (2009) in our experiments, since theirs are well-known and standard methods.

Banea's method (2008) has two classification models that are dependent on the running position of the MT system.

Training data Translation Model (TrTM) This model is actually shown in Figure 2. A text classifier is learned using a dataset described in the target language. To do so, before the text classifier is learned, documents (reviews) in the training dataset that are described in the source language are translated into the same language as those in the test dataset. We do not need to do anything with the test dataset.

Test data Translation Model (TeTM) This is a reverse version of TrTM. A text classifier is learned using a dataset described in the source language. In this model, documents in the test dataset are translated into the same language as that in the training dataset before the classification phase is run. We do not need to do anything with the training dataset.

Wan's method (2009) combines the above two models through the multi-viewpoint style co-training approach proposed by Blum and Mitchell (1998). Here, the source language and the target language are considered as each viewpoint. The sets of features extracted from dataset described in each language are simultaneously used in the co-training framework. This method iteratively runs TrTM and TeTM. For each iteration, two sets of additional unlabeled review dataset, one is described in the target language and another is the same dataset but is translated into the source language, are applied as input to TrTM/TeTM to predict their (temporal) class label. Of all predicted review data, a subset confidently predicted is added into the original labeled training dataset. We call this method the **Co-training Model** in the remainder of this paper.

The sentence filtering mentioned in the previous section is a preprocessing stage of multilingual review classification. Therefore, each classification model (TrTM, TeTM, and Co-training) is able to run without any modifications. We can directly use the snippets as elements of the training/test dataset.

4.1.2 Dataset

Works on sentiment analysis have usually been carried out in English because there is a large amount of English linguistic resources available for sentiment analysis. Thus, in this study we set English as a source language and Japanese as a target language.

We collected reviews for use in our experiments from one of the most popular global e-commerce sites, Amazon. We accessed Amazon.com ("http://www.amazon.com/") for English reviews and Amazon.co.jp ("http://www.amazon.co.jp/") for Japanese reviews.

Table 1: Number of English/Japanese polarity words

polarity words	all	positive	negative
English	1,392	609	783
Japanese	724	340	384

Table 2: Number of documents/sentences including a polarity word

data type	#documents		#sentences	
English	9,738/10,000	(97%)	51,661/82,310	(63%)
EtoJ	8,283/10,000	(83%)	26,424/82,310	(32%)
Japanese	955/ 1,000	(96%)	3,498/ 7,466	(47%)
JtoE	985/ 1,000	(99%)	5,017/ 7,466	(67%)

First, we prepared a common product list. This is a list of products that can be purchased through both Amazon.com and Amazon.co.jp. We used in this study a list of MP3 audio players, such as "iPod (Apple)" and "Walkman (Sony)". Second, we retrieved and crawled a set of reviews by using the above list from Amazon.com and Amazon.co.jp. All crawled reviews hold an up-to-five-star user rating. We regarded reviews holding four or five stars as positive reviews and those holding one or two stars as negative reviews. As a result, we obtained 1,000 Japanese reviews (500 positive / 500 negative reviews), and 10,000 English reviews (5,000 positive / 5,000 negative reviews). In our setting, the source language was English. The volume of English reviews was 10 times that of Japanese ones. All reviews were original, and there were no duplicates.

4.1.3 Polarity dictionary

We need to prepare a set of polarity words to run sentiment-oriented sentence filtering. We used a polarity dictionary generated as follows.

1) We constructed initial polarity dictionaries by using the methods by Takamura et al. (2005b) and Takamura et al. (2005a) [2]. In these methods, the English polarity dictionary is constructed based on WordNet (1998) information, and the Japanese polarity dictionary is constructed based on Iwanami Japanese-language dictionary (1994), respectively. Each method output a set of word/polarity pairs with a confidence level.

2) We manually corrected words with a high confidence level, and we eliminated words with a low confidence level from the initial dictionary.

Table 1 shows the number of English/Japanese polarity words in our dictionary.

Table 2 shows the number of documents/sentences including a polarity word in the dataset. The abbreviation EtoJ means English documents were translated to Japanese. The abbreviation JtoE means translation in the opposite direction. On the document level, excepting the case of EtoJ (83%, slightly low percentage), almost all documents (reviews) included at least one polarity word. This means that the set of polarity words used in the experiments has wide coverage.

4.1.4 Other settings

We used as a machine translation system the Excite automatic translation service[3]. This site provides rule-based machine translation between English and Japanese (both EtoJ and JtoE).

For learning review classifiers, we used a linear kernel support vector machine (SVM) and the software package Classias[4] for training SVM classifiers. Unigram-based binary feature vectors were constructed. As the tokenization process (recognizing word separations) for Japanese reviews, we used a well-known Japanese NLP programming software package, MeCab[5]. All English words in

[2]The essential part of the above both papers is the same. The difference is only that language for the input. In the (Takamura et al., 2005b) the authors introduced for the English polarity dictionary, and in the (Takamura et al., 2005a) introduced for the Japanese polarity dictionary.

[3]http://www.excite.co.jp/world/

[4]http://www.chokkan.org/software/classias/index.html.en

[5]http://mecab.sourceforge.net/

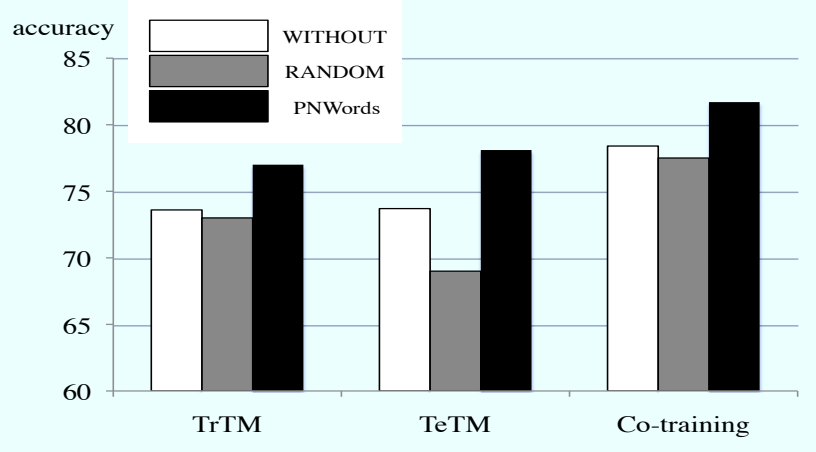

Figure 5: Effects of sentence extraction

the dataset were lower-cased.

We used ten-fold cross-validation for the evaluation.

4.2 Experimental results

The experimental results are shown in Table 3 (see also Figure 5). The value in each cell indicates the classification accuracy. Each column shows the multilingual review classification method, and each row shows the sentence extraction method in the sentence filtering step. **PNWords** is the sentence extraction method described in Section 3, i.e., our proposed method. The others are baseline methods for comparison. **WITHOUT** means that the sentence filtering step was skipped at the training phase of text classifiers; all sentences in the reviews in the training dataset were used in the training phase. **RANDOM** means that snippets were generated by randomly extracting K percent of sentences from the original reviews in the dataset. We set K=50 in the experiments. Unlike WITHOUT and PNWords, RANDOM had essentially randomness. Therefore, we prepared five sets of snippets by running RANDOM five times and then measured five accuracy values. The average accuracy is shown in Table 3.

We also developed a system which was trained on documents written in Japanese in order to see what is the accuracy of the system when a MT is not used. The accuracy of this system is 77.9%.

To investigate the performances of the three multilingual classification methods, we first ignored the effects of sentence filtering modules and simply compared the accuracies of the first row, i.e., the results obtained by WITHOUT. Table 3 shows that the accuracy of Co-training is higher than that of both TrTM and TeTM. Thus, the co-training model is considered to have an advantage over both TrTM and TeTM. This result corresponds with those reported by Wan (2009). We confirmed that Wan's co-training method outperforms TrTM and TeTM in a multilingual review classification problem.

Table 3: Effects of sentence extraction

	TrTM	TeTM	Co-training
WITHOUT	73.6	73.7	78.4
RANDOM	73.0	69.0	77.5
PNWords	77.0	78.1	**81.7**

Next, we investigated the effectiveness of the proposed sentence filtering method. In comparing WITHOUT and RANDOM for each multilingual review classification method, when the sentence filtering step with the RANDOM method was added to the training phase of text classifiers, the classification accuracy worsened rather than improved. One can see that extracting sentences without thought (namely, at random) does not contribute to improvement of the text classification performance. Last, in comparing WITHOUT and PNWords, one can see that PNWords outperforms WITHOUT for all the multilingual review classification methods and that the combination of Co-training and PNWords achieves the best performance. From the results, we can conclude that our sentiment-oriented sentence filtering method can improve multilingual review classification.

5 Related Works

Several methods of monolingual document-level sentiment classification have been proposed. In the early works in this field, such as by Pang et al. (2002), Mullen and Collier (2004), and Gamon (2004), the interest was in simply applying machine learning approaches. The latest works in this field have discussed some specific features for sentiment analysis. For example, Li et al. (2009) and Dasgupta and Ng (2010) considered shifting polarity and ambiguous polarity in documents.

The multilingual setting is also a recent topic. As described in Section 4, Banea et al. (2008) proposed a simple solution using machine translation. Wan (2009) extended Banea's work, and applied for English/Chinese reviews. Denecke (2008) also proposed a similar method for English/German reviews. He used SentiWordNet[6], which is an enhanced lexical resource for sentiment analysis and opinion mining.

In the word-level multilingual sentiment classification area, Mihalcea et al. (2007) proposed two methods for translating polarity words using bilingual dictionaries and a parallel corpus. Scheible (2010) proposed a graph-based approach to obtain translation information of polarity words. He used English/German dataset.

In the sentence-level multilingual sentiment classification area, Banea et al. (2010) conducted experiments with six languages (English, Arabic, French, German, Romanian and Spanish) , and reported that one can predict sentence-level subjectivity in languages other than English, by leveraging on a manually annotated English dataset, with 71.3% (for Arabic) to 73.66% (for Spanish).

6 Conclusion

We investigated the effectiveness of applying our sentiment-oriented sentence filtering method to reduce the influence of translation errors in multilingual document-level review classification. Our filtering method can improve the performance of multilingual review classification. Experimental results showed that the proposed method achieved 81.7% classification accuracy.

The following issues will need to be addressed to refine our method.

- In this study, we treated sentence-level linguistic units to reduce the influence of translation errors. In the future, we will also investigate performances when extracting fine-grained linguistic units, such as words and phrases. For example, Wei and Pal (2010) attempted to apply structural correspondence learning (Blitzer et al., 2006; Blitzer et al., 2007) to find a low dimensional document representation.

- We applied the proposed method only to English/Japanese dataset. Additional experiments with other languages should be conducted for further and more sophisticated data analysis.

- Yang et al. (2009) handled heterogeneous data in a framework of transfer learning (Pan and Yang, 2010). The relationship between our approach and transfer learning would be interesting to examine.

[6] http://sentiwordnet.isti.cnr.it/

References

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of Recent Advances in Natural Language Processing*.

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

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 28–36.

John Blitzer, Ryan McDonald, and Rernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 120–128.

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 for Computational Linguistics*, pages 440–447.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Sajib Dasgupta and Vincent Ng. 2010. Mine the easy and classify the hard: Experiments with automatic sentiment classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 701–709.

Kerstin Denecke. 2008. Using SentiWordNet for multilingual sentiment analysis. In *Proceedings of the ICDE Workshop on Data Engineering for Blogs, Social Media, and Web 2.0*, pages 507–512.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Michael Gamon. 2004. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 18th International Conference on Computational Linguistics*.

Soo-Min Kim and Eduard Hovy. 2005. Automatic detection of opinion bearing words and sentences. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 61–66.

Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu-Ren Huang, and Guodong Zhou. 2009. Sentiment classification and polarity shifting. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

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 for Computational Linguistics*.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 412–418.

M. Nishio, E. Iwabuchi, and S. Mizutani. 1994. *Iwanami Japanese-language dictionary*. Iwanami Shoten.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, pages 271–278.

Bo Pang and Lillian Lee. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 76–86.

Christian Scheible. 2010. Sentiment translation through lexicon induction. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 25–30.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005a. Extracting semantic orientation of words using spin model. In *IPSJ SIG Note (NL-168-22)*, pages 141–148. (In Japanese).

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005b. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 133–140.

Huifeng Tang, Songbo Tan, and Xueqi Cheng. 2009. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7):10760–10773.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 235–243.

Bin Wei and Christopher Pal. 2010. Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 258–262.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the ACM 14th Conference on Information and Knowledge Management*.

Qiang Yang, Yuqiang Chen, Gui rong Xue, Wenyuan Dai, and Yong Yu. 2009. Heterogeneous transfer learning for image clustering via the social web. In *Proceedings of the ACL-IJCNLP*, pages 1–9.