CROSS-LINGUAL SENTIMENT LEXICON LEARNING

GAO DEHONG

Ph.D

The Hong Kong Polytechnic University

2014

The Hong Kong Polytechnic University

Department of Computing

Cross-lingual Sentiment Lexicon Learning

GAO Dehong

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

February 2014

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

__________________________ (Signed)

GAO Dehong (Name of Student)

Abstract

Sentiment lexicon contains a certain number of known-sentiment words (e.g., "*good*", "*nice*" and "*bad*"). It has been widely recognized that sentiment lexicon plays a fundamental role in sentiment analysis. Relative to the existing sentiment lexicons in English, the available sentiment lexicons in the other languages such as Chinese are far from sufficient. This dissertation focuses on **C**ross-lingual **S**entiment **L**exicon **L**earning (CSLL), whose goal is to make full use of the existing sentiment resources from one (or more) language(s) to automatically learn sentiment lexicon(s) for other language(s).

The dissertation work makes a systematic study on CSLL. In *bilingual graph based sentiment lexicon learning*, a bilingual graph is built with the words in English and in a target language for which we want to generate the sentiment lexicon. A label propagation based approach is proposed to transfer the sentiment information from English to the target language. To the best of our knowledge, the word alignment information derived from the parallel corpus is the first time leveraged to build the inter-language relations in CSLL, which is proved to significantly increase the coverage of the learned sentiment lexicon. In this work, the sentiment polarity of a word is determined by the sentiment information of the connected words in the bilingual graph. In *Co-training based bilingual sentiment lexicon learning*, we consider not only the sentiment information of the connected words, but also the information about the words themselves (e.g., word definitions). From these two types of information, novel and effective features are explored to deduce the sentiment polarity of a word. With these

features, CSLL is considered as word level sentiment classification and the two classifiers are developed based on the co-training framework to predict the sentiment polarities of the words in two languages respectively. In particular, the learning processes of the two classifiers are connected by the word associations derived from the bilingual resources (e.g. bilingual dictionaries). In these two pieces of work, the words with similar semantics are assumed to have similar sentiments. The proposed approaches can thus connect or associate the semantic-similar words in the learning processes. However, the words similar in semantics do not always have the similar sentiments, especially when the words have multiple senses. In *multilingual sentiment lexicon learning*, we are dedicated to automatically refine the semantic-oriented connections to the sentiment-oriented connections. Incorporating with multilingual (sentiment) resources, a novel label propagation based approach is developed to propagate sentiment information between multiple languages and to automatically update the weights of the connections. The main contribution of this work is that the proposed approach not only performs well in multilingual sentiment lexicon learning, but also provides a new strategy for graph update. Extensive experiments have been conducted in each piece of work and experimental results demonstrate the effectiveness of the approaches proposed.

To summarize, as one of the few large-scale studies on CSLL, this dissertation provides complete learning techniques and a deep analysis on the key factors for cross-lingual sentiment lexicon learning.

Publications Arising from the Thesis

1. **Dehong Gao**, Furu Wei, Wenjie Li, Xiaohua Liu and Ming Zhou. Cross-lingual Sentiment Lexicon Learning with Bilingual Graph Label Propagation. *Computational Linguistics*. 2014.
2. **Dehong Gao**, Wenjie Li, Renxian Zhang, Xiaoyan Cai and You Ouyang. Sequential Summarization: a Full View of the Twitter Trending Topics. *IEEE Transactions on Audio, Speech & Language Processing*. 2013.
3. **Dehong Gao**, Furu Wei, Wenjie Li and Ming Zhou. Sentiment-orientated Label Propagation: Identify Chinese Word Sentiment with Multilingual Resources. *ACM Transactions on Asian Language Information Processing,* Article under review.
4. **Dehong Gao**, Wenjie Li and Renxian Zhang. Sequential Summarization: a New Application for Timely Updated Twitter Trending Topics. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL'13), 2013.
5. **Dehong Gao**, Furu Wei, Wenjie Li, Xiaohua Liu and Ming Zhou. Co-training Based Bilingual Sentiment Lexicon Learning. *In Proceedings of the 27th AAAI Conference on Artificial Intelligence Late Breaking Paper* (AAAI'13 LBP), 2013.
6. Renxian Zhang, Wenjie Li and **Dehong Gao**. Towards content-level coherence with aspect-guided summarization. *ACM Transactions on Speech and Language Processing*. 2013.
7. Renxian Zhang, Wenjie Li, **Dehong Gao** and Ouyang You. Automatic

Twitter Topic Summarization with Speech Acts. *IEEE Transactions on Audio, Speech & Language Processing*. 2013.

8. Fanqi Meng, **Dehong Gao**, Wenjie Li, Xu Sun and Yuexian Hou. A Unified Graph Model for Personalized Query-Oriented Reference Paper Recommendation. *In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (CIKM'13), 2013.
9. **Dehong Gao** and Wenjie Li. Opinion Influence and Diffusion in Social Network. *Doctorial Consortium of ACM's Special Interest Group on Information Retrieval* (SIGIR'12 DC). 2012.
10. **Dehong Gao**, Renxian Zhang, Wenjie Li and Yuexian Hou. Twitter Hyperlink Recommendation with User-Tweet-Hyperlink Three-way Clustering. *In Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (CIKM'12), 2012.
11. **Dehong Gao**, Wenjie Li and Renxian Zhang. Beyond Twitter Text: A preliminary Study on Twitter Hyperlink and its Application. *In Proceedings of the 24th International Conference on Computational Linguistics* (Coling'12), 2012.
12. **Dehong Gao**, Wenjie Li and Renxian Zhang. Efficient Feedback-based Feature Learning for Blog Distillation as a Terabyte Challenge. *In Proceedings of the 24th International Conference on Computational Linguistics* (Coling'12), 2012.
13. **Dehong Gao**, Wenjie Li and You Ouyang. LDA-based Topic Formation and Topic-Sentence Reinforcement for Graph-Based Multi-Document Summarization. *In Proceedings of the 9th Asia Information Retrieval Societies Conference* (AIRS'12), 2012.

14. Renxian Zhang, Wenjie Li and **Dehong Gao**. Generating Coherent Summaries with Textual Aspects. *In Proceedings of Association for the Advancement of Artificial Intelligence* (AAAI'12), 2012.
15. Renxian Zhang, **Dehong Gao** and Wenjie Li. Towards Scalable Speech Act Recognition in Twitter: Tackling Insufficient Training Data. *In Proceedings of EACL'12 Workshop on Semantic Analysis in Social Networks*, 2012.
16. Junyi Chai, James N.K. Liu, **Dehong Gao** and Jian Xu. A new tolerant skyline operator for decision support. *In Proceedings of IEEE International Conference on Granular Computing* (GrC'12), 2012.
17. **Dehong Gao**, Renxian Zhang, Wenjie Li, Raymond Yiu-Keung Lau and Kam-Fai Wong. Learning features through feedback for blog distillation. *In Proceedings of ACM's Special Interest Group on Information Retrieval* (SIGIR'11), 2011.
18. Renxian Zhang, **Dehong Gao**, and Wenjie Li. What Are Tweets Doing: Recognizing Speech Acts in Twitter. *In Proceedings of AAAI'11 Workshop on Analyzing Microtext*. 2011.
19. Xiaoyan Cai, Renxian Zhang and **Dehong Gao**. Simultaneous Clustering and Noise Detection for Theme-based Summarization. *In Proceedings of the 5th International Joint Conference on Natural Language Processing*. 2011.
20. Wenping Zhang, **Dehong Gao** and Wenjie Li. PolyU at MOAT in NTCIR-8. *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*. (NTCIR8) 2010.

Acknowledgements

Having completed this doctoral dissertation, I hope to give my sincere appreciation to all those who offered help. My first thanks go to Prof. Wenjie Li, my supervisor who gives me the valuable assistance throughout my whole study in the Hong Kong Polytechnic University. Her attitudes to the academic research and to the real lift have guided me towards both doing better research and being a better person. Without her consistent and illuminating instructions, this dissertation could not have reached its present appearance.

Please allow me to express heartfelt appreciation to Prof. Kam Fai Wong, my co-supervisor, for his support and trust in the past four years. Furthermore, I am also grateful to Dr. Yuexian Hou, Prof. Qin Lu, Prof. Jane You and Dr. Lau Yiu Keung, for their instructions and suggestions during my study. Specially, I would like to thank Dr. Yan Liu for her enlightening comments and inspiring advices in my confirmation and studies.

At the same time, I owe my cordial appreciation to Dr. Furu Wei, my mentor at Microsoft Research. Our discussion greatly underpins my enthusiasm for the subject of cross-lingual sentiment lexicon learning. As well I am grateful for all the generous help from Dr. Ming Zhou, Dr. Xiaohua Liu, and Dr. Wei Wu during my internship at Microsoft Research. This period of research in Microsoft Research makes preparation for this dissertation.

My thanks are also given to my best friends Dr. Xiaoyan Cai, Dr. Libin Yang, Dr. Renxian Zhang, Dr. Ouyang You, Dr. Junyi Chai, Mr. Jian Xu, Mr. Dong Liang, Ms. Yanning Wang, Mr. Zhouxiang Zhang, Ms. Dan Xiong, Ms. Chengyao Chen and etc., who lighten my life in Hong Kong.

Finally, I owe a debt of gratitude to my beloved parents and girlfriend for having supported me with confidence and tender care through all these years. Their fountain love has always been watering my mind and warming my heart. I sometimes felt guilty of not being able to spend much time with them but their comprehension and concern helped me get through the hardest days in my life. So I hope to use this precious opportunity to express thanks to my family, as well as my bigger family of friends and teachers. You are my most precious in life.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

[1]

Sentiment analysis, as a leading-edge area in Natural Language Processing (NLP), has been actively pursued by world-wide researchers since 2000 [Pang and Lee 2008; Liu 2012]. One of the fundamental issues in sentiment analysis research is automatic sentiment lexicon learning [Qiu et al. 2011]. The sentiment lexicon contains certain amounts of known-sentiment words, such as “good”, “beautiful”, “bad”, and “terrible”, which act as the indicators of sentiment polarities of the text units (e.g., documents and sentences). The sentiment lexicons, e.g., the Harvard General Inquirer (GI) lexicon [Stone et al. 1963] used in earlier studies are mostly compiled by human experts. To avoid the labor-intensive and time-consuming annotation, researchers have presented fruitful studies on automatically sentiment lexicon learning. However, most prior work focuses on English sentiment lexicon learning or expansion, while the work on other languages is not well established. For example, the available Chinese sentiment lexicons are rare for the public research, and even if they exist, they

[1] The figure is created by Wordle (www.wordle.net) with the input data collected from the GI lexicon.

are far behind English in both quantity and quality. To alleviate this problem, I have been devoted to cross-lingual sentiment lexicon learning which aims to take advantage of existing sentiment resources from one (or more) language(s) to build sentiment lexicon(s) for other language(s). The approaches developed in this dissertation are all language-independent and thus they can be applied to learn the sentiment lexicons for any other languages.

Given that a sentiment lexicon plays a fundamental role in sentiment analysis, I will first give an overview of sentiment analysis research. And then I will explain the motivation for cross-lingual sentiment lexicon learning, summarize the contributions of my work and describe the overall structure of the dissertation.

1.1 Sentiment Analysis: a Brief Introduction

Sentiment analysis is defined as a study to analyze people's opinions, sentiments, appraisals and attitudes towards the objects such as products, services, institutions and their specific aspects [Liu 2012]. It is one of the most popular research areas in natural language processing. The mushrooming of researches on sentiment analysis has been witnessed since the year of 2000. The developed techniques have been already applied to many applications, such as, political election prediction [Tumasjan et al. 2010; Yano and Smith 2010; Chen et al. 2010], stock market analysis [Bollen et al. 2007; Feldman et al. 2011] and etc.

Given a text unit (a document, a paragraph, or a sentence), the objective of sentiment analysis is to determine whether this text unit contains an opinion, and then to discover who (*opinion holder*) have what sentiment (expressed by *sentiment word*) on which attribute (*aspect*) of which object (*opinion target*) at which time (*time*). The five composes of a complete opinion expression [Liu 2012]. The main studies in sentiment analysis are conducted along four directions with reference to their expected research achievements.

- *Sentiment classification*: Sentiment classification determines whether a text unit contains an opinion and identifies the sentiment polarity (e.g., positive or negative) delivered in the text unit. It has been widely recognized by the researchers as one of the most extensively studied tasks in sentiment analysis.
- *Sentiment extraction*: Sentiment extraction recognizes the opinion expressions in the text units and extracts one or more components (e.g., opinion target, its aspects, sentiment words etc.) from the expressions. Existing work mainly focuses on aspect extraction, for example identifying the product aspects from the product reviews.
- *Opinion summarization*: While the above two normally analyze or extract a single piece of opinion expressed in the text unit, opinion summarization, on the other hand, focuses on automatically summarizing the sentiment information discovered in a collection of sentiment-bearing documents.
- *Opinion influence and diffusion*: Benefiting from Web2.0 techniques, the increasing social network applications, like Twitter (www.twitter.com) and Facebook (www.facebook.com) provide both new opportunities and new challenges to sentiment analysis. Opinion influence and diffusion explore how opinions propagate over social networks, including the formation of the public opinions, the opinion influence among individual users or groups of users and the opinion

dissemination via social network connections [Gao 2012].

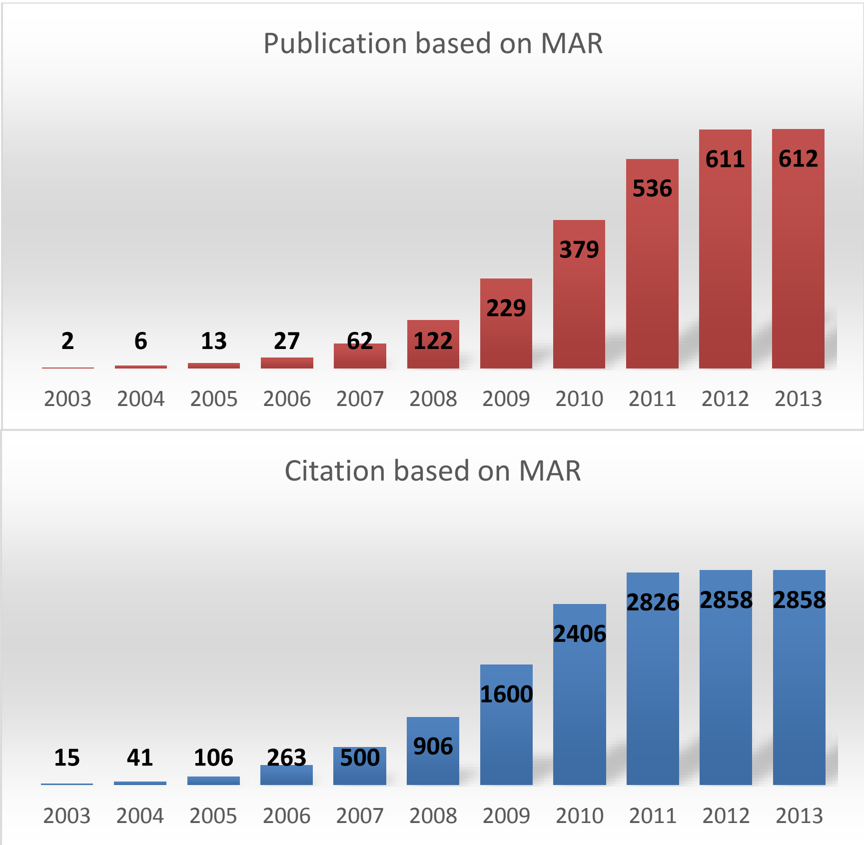


Figure 1.1 Numbers of publications and citations on "sentiment analysis" in the passing ten years

To show the progress of the sentiment related research, the numbers of publications and citations[2] on sentiment analysis in the past ten years are plotted in the above Figure 1.1. It is obvious that the research on sentiment analysis is increasing significantly these years. Meanwhile, a development line of sentiment analysis is observed by reading these articles. The research targets in these articles gradually shift from documents or sentences to phrases or words. Most of early researches on sentiment analysis are conducted on document level or

[2] These survey results are generated from Microsoft Academic Research (MAR) (academic.research.microsoft.com) on 25 September 2013 by taking the phrase "sentiment analysis" as the search query.

sentence level sentiment classification, which focuses on the overall sentiment polarities conveyed in the documents or sentences. Later on, increasing studies emerge on sentiment extraction, which focuses on sentiment analysis based on the opinionated phrases. Recent years, the studies on sentiment-oriented lexical resources keep rising [Anbananthen and Elyasir 2013]. One reason for this development is that the practical applications are not satisfied with the systems that can only provide the overall sentiment polarities of the documents or sentences. They require the systems to deeply understand the documents or sentences and to explore more delicate sentiment information from the phrases or words. Another reason is that more and more researchers begin to realize the important role of the basic sentiment resources (e.g., sentiment lexicons) in sentiment analysis. Thus, the research targets gradually shift from document or sentence level sentiment analysis to phrase or word level sentiment analysis.

1.2 Research Motivation

It has been widely recognized that sentiment lexicon is the most valuable resource in the community of sentiment analysis [Pang and Lee 2008; Liu 2012]. Most of the above mentioned research studies are related to sentiment lexicons explicitly or implicitly.

As reported in [Qiu et al. 2011], sentiment lexicon plays an important role in sentiment classification. Generally, the sentiment classification approaches are divided into unsupervised approaches [Turney 2002; Kim and Hovy 2004] and (semi-)supervised approaches [Pang et al. 2002; Zhang et al. 2010; Abbasi et al. 2008]. For unsupervised sentiment classification approaches, the sentiment information of words (or phrases) is the basis which needs to be identified beforehand. The sentiment polarity of a text unit is then determined by aggregating the sentiment information from all the words (or phrases) in the text unit. For (semi-)supervised approaches, the sentiment words can be used as the classification features [Kim and Hovy 2006]. It is found in [Zhang et al. 2010; Gao et al. 2011] that the sentiment words are effective features in sentiment classification. Sentiment lexicon is also related to sentiment extraction. Intuitively, the sentiment components (i.e., sentiment words, opinion targets, aspects, opinion holders and time) are concurrent in an opinionated text unit [Brody and Elhadad 2010]. If the sentiment words are able to be identified by sentiment lexicons, the rest sentiment components can be also extracted by manually-defined syntactic rules [Ku et al. 2006]. Likewise, the other work in sentiment analysis is all, more or less, related to sentiment lexicons.

The sentiment lexicons, such as the Harvard General Inquirer (GI) lexicon and Micro-WNOp[3] [Wilson et al. 2005; Wiebe et al. 2005] are often created by experts. However, manually compiling a sentiment lexicon is labor-intensive and time-consuming. Though the human-compiled sentiment lexicons are accurate in word sentiment polarity, the compiled lexicons are usually low in coverage as people express sentiments colloquially and the sentiment words evolve along time frequently. In consideration of this, researchers conduct fruitful studies to learn sentiment lexicons automatically. By now, most of these studies focus on English sentiment lexicon induction and expansion, while the study on other languages is still in the early stages.

To this end, this dissertation is devoted to the issue of *cross-lingual*

---

[3] http://www-3.unipv.it/wnop

sentiment lexicon learning, i.e., given the sentiment resources in one (or more) language(s), it automatically learns the sentiment lexicons for other language(s). The aim is to develop the generic approaches, which are language-independent and can be applied to sentiment lexicon learning in any language. The generated sentiment lexicon is expected to provide general sentiment information (e.g., both positive and negative scores) for a word, like *SentiWordNet* [Baccianella et al. 2010]. The public users are allowed to use these sentiment scores in their applications and adjust based on their own contexts.

1.3 Research Overview and Contributions

In this dissertation, a systematical study is conducted on cross-lingual sentiment lexicon learning from three perspectives. The two main issues in CSLL are *word sentiment representation* (i.e., how to select sentiment-related features) and *cross-lingual sentiment information mapping* (i.e., how to transfer sentiment information between languages). We first leverage bilingual graph to model these two parts in *bilingual graph based sentiment lexicon learning* (Chapter 3). Then, we improve word sentiment representation and cross-lingual sentiment information mapping in *co-training based bilingual sentiment lexicon learning* (Chapter 4) and *multilingual sentiment lexicon learning* (Chapter 5), respectively.

In bilingual graph based sentiment lexicon learning, a bilingual graph is established with the words in English and in a target language for which we want to generate the sentiment lexicon. Specifically, the words in the two languages are bridged with the inter-language relations and the words in the same languages are connected with the intra-language relations. Based on the bilingual graph, the label propagation based approach is proposed to induce the sentiment lexicon for the target language with the help of English sentiment seed words. To the best of our knowledge, this work is the first one to leverage the word alignment information derived from the parallel corpus to build the inter-language relations in cross-lingual sentiment lexicon learning. The experimental results demonstrate the advantage of the use of word alignment in building the inter-language relations.

In bilingual graph based sentiment lexicon learning, the sentiment polarity of a word is determined by the sentiment information of the connected words and by the relations (e.g., synonyms or antonyms) between the connected words. The information related to the word itself (e.g., word definition or explanation), however, is not taken into consideration. From these two types of information, numbers of features can be explored to indicate word sentiment polarity. The features can be categorized into the text-related features and the graph-related features. With these features, sentiment lexicon learning can then be treated as word level sentiment classification. A bilingual sentiment lexicon learning approach based on the co-training framework is proposed to simultaneously generate sentiment lexicons for the two languages with the available sentiment information. For each language, a classifier is developed to effectively integrate the text-related and graph-related features. In each iteration of learning, if the classifier can confidently predict the sentiment polarity of an unlabeled word, the words associated with the unlabeled word in the other language by bilingual resources (e.g., bilingual dictionaries) are also assigned with the same sentiment polarity. These words with their sentiment polarities are then used to update the training datasets of the corresponding languages. After co-training, the classifier is then used to generate the sentiment lexicon by predicting the sentiment polarities of unlabeled words. Extensive experiments have been conducted to verify the roles of the features in bilingual sentiment lexicon learning.

The assumption made in the previous work is that words similar in semantics are also similar in sentiment. Based on this assumption, the semantics-similar words are connected in learning. Actually, these connections cannot guarantee that the connected words are always similar in sentiment polarity, especially when a word has multiple meanings. For example, the English word “blue” and the Chinese word “蓝色” (of the color blue, a non-sentiment word) share a common meaning. Nevertheless, “Blue” can be a negative word which means low in spirits in certain context in English, while “蓝色” is always a non-sentiment bearing word in Chinese. In multilingual sentiment lexicon learning, a sentiment-oriented label propagation approach is proposed to propagate the sentiment information among multiple languages and simultaneously to increase (or decrease) the weights of the connections that are (not) likely to connect the words with the similar sentiment polarities. The sentiment words from multiple languages are introduced in this work not only to bring in more sentiment information, but also to guide the weight update of the connections. The experimental results show the advantages of the proposed approach in multilingual sentiment lexicon learning. The main contribution of this work is the novel idea of weight update, which can be also applied to the other problems where the specific connections are insufficient or hard to acquire. The proposed approach allows people to start with collecting as many connections as possible, and then to refine those connections for the specific problems.

1.4 Structure of Dissertation

The overall picture of the dissertation is as follows.

Chapter 1 briefly describes the major studies on sentiment analysis with special focus on the role of sentiment lexicon learning in sentiment analysis. The main motivation and the structure of this dissertation are presented.

Chapter 2 surveys the existing work on sentiment classification, sentiment extraction and opinion summarization. After presenting the literature on these areas, the focus then shifts to the studies devoted to sentiment lexicon learning.

Chapter 3 is engaged to bilingual graph based sentiment lexicon learning, where the label propagation based approach to induce the sentiment lexicon for a given language is presented.

Chapter 4 discusses co-training based bilingual sentiment lexicon learning from the view of word level sentiment classification, where two classifiers are developed to learn the sentiment lexicons for the two languages simultaneously.

Chapter 5 is devoted to multilingual sentiment lexicon learning, where the proposed approach transfers sentiment information among multiple languages and leverages the multilingual sentiment information to automatically refine word connections.

Chapter 6 summarizes the major findings, conclusions and contributions of the work. At last, the future directions of cross-lingual sentiment lexicon learning are pinpointed.

Chapter 2 Literature Review

Sentiment analysis, as one of the most frontier tasks in natural language processing, has attracted considerable attention from its outset in the late 1990s. Worldwide researchers have contributed fruitful ideas and techniques to sentiment analysis research, including sentiment classification, sentiment extraction, opinion summarization and sentiment lexicon learning. This chapter introduces the overall picture of these research studies and reviews the related work. In particular, Section 2.1 surveys the main research directions in sentiment analysis. Section 2.2 reviews the approaches specially proposed for sentiment lexicon learning. Finally, Section 2.3 concludes this chapter.

2.1 Research in Sentiment Analysis

As mentioned in Chapter 1, there are four main lines of research in sentiment analysis, among which sentiment classification, sentiment extraction and opinion summarization are closely related to sentiment lexicon learning either explicitly or implicitly. The following sections introduce the core techniques applied in the existing work.

2.1.1 Sentiment Classification

Sentiment classification[4] is the most extensively studied topic in sentiment analysis. It aims to classify text units into special sentiment polarities (e.g., positive or negative) [Liu 2012]. There are several ways to categorize these studies by different criteria. For example, according to the text unit to be classified, the studies can be divided into document-level or sentence-level sentiment classification, where a document or a sentence is regarded as the basic information unit and it conveys only one sentiment polarity as a whole. However, in this survey, I follow [Pang and Lee 2008]'s way to categorize these studies into unsupervised, supervised and semi-supervised sentiment classification according to the classification techniques they use. With such categorization, the role of sentiment lexicon learning can be explained more clearly.

2.1.1.1 Unsupervised Sentiment Classification

In general, an *unsupervised* approach determines the sentiment polarities of the opinion indicators (usually words or phrases) in a text unit, either a document or a sentence. The overall sentiment polarity of the whole text unit is then calculated by aggregating the sentiment polarities of all its indicators. The sentiment polarity information of the indicators is the prior knowledge and forms the basis of the unsupervised approaches.

Most early studies on sentiment classification fall into this category. For example, [Turney 2002] presents an unsupervised learning approach that

[4] According to [Pang and Lee 2008], sentiment regression and ranking are all categorized into sentiment classification.

involves the following three steps. (1) Frequent phrases are extracted according to Part-Of-Speech (POS) tags and manually-defined patterns (e.g., the first word is an adjective and the second is a noun or a noun phrase). (2) The sentiment polarities (i.e., positive or negative) of the extracted phrases are determined by the Point-wise Mutual Information (PMI) [Fano 1961] with two standard positive and negative words, i.e., "excellent" and "poor". (3) The sentiment polarity of a sentence is computed by the average sentiment polarities of all the extracted phrases in the sentence. The same approach is used in [Li et al. 2011], but it studies sentiment classification of Chinese reviews. Rather than using two reference words only, the approach in [Yu and Hatzivassiloglou 2003] works with two sets (i.e., positive and negative) of sentiment seed words and with a newly-defined semantic orientation measurement.

[Hu and Liu 2004] generates a domain-specific sentiment lexicon by using certain sentiment seed words and their synonyms and antonyms found in WordNet. This work also considers the effect of negation words (e.g., "no" and "not") and contrary words (e.g., "but" and "however") when determining the sentiment polarity of the text unit. Like in the previous work, the sentiment polarity of a product review is determined by the sentiment polarities of all the sentiment words in the review in [Kim and Hovy 2004]. However, instead of summing up the sentiment scores of the words, their work calculates the sentiment polarity of a review by multiplying the sentiment scores of each sentiment word in the review.

The above mentioned work all regards the sentiment polarity determination of the indicators as a subtask of sentiment classification of the text unit. Alternatively, some other unsupervised approaches simply rely on the sentiment words in the existing sentiment lexicons. In contrast, they mainly focus on developing more sophisticated algorithms to aggregate the sentiment polarities of the indicators [Nasukawa and Yi 2003; Ding et al. 2008; Taboada et al. 2011].

2.1.1.2 Supervised Sentiment Classification

Sentiment classification can be treated as a kind of text categorization [Aggarwal and Zhai 2012]. The main difference between sentiment classification and general text categorization is that the classes of the former are sentiment-oriented classes, like positive *v.s.* negative, 1~5 stars and etc., while the latter classifies the texts into the classes of topics such as politics, sports, and technology etc. Like in text categorization, machine learning approaches are also most widely used in sentiment classification. Many current work have been proposed to explore machine learning approaches. They fall into *supervised sentiment classification*.

The supervised sentiment classification approach is first introduced by [Pang et al. 2002], which aims to classify the movie reviews as either positive or negative. Three machine-learning approaches, namely Naïve Bayes (NB) classifier [Zhang 2004], Maximum Entropy (ME) classifier [Nigam 1999] and Support Vector Machine (SVM) classifier [Cortes and Vapnik 1995; Joachims 1999] are employed. These classifiers are trained based on the *N*-gram features. The major conclusions from this paper are: (1) unigram is more effective than bigram; (2) a better performance is achieved by the presence rather than the frequency of the features and (3) both the NB classifier and SVM classifier show their advantages in classification.

Later on, richer features and more sophisticated learning algorithms are further explored in the subsequent research work. With regard to the classification features, [Dave et al. 2003] use the bigram and trigram features extracted from the product reviews. The experiment results show that based on their dataset bigrams and trigrams features are more powerful in classification. This does not comply with the first conclusion from [Pang et al. 2002]. The other features that can be used in sentiment classification include positional features and opinion-bearing words [Kim and Hovy 2006], the Part-Of-Speech information [Mullen and Collier 2004; Whitelaw et al. 2005], negation words [Na et al. 2004; Kennedy and Inkpen 2006], contextual valences and sentiment shifters [Kennedy and Inkpen 2006; Li et al. 2010b]. The topic-related features are also shown to play a critical role in sentiment classification [Hu and Li 2011; Gao et al. 2012b]. I have also conducted experiments on sentiment classification and find that the sentiment words are also effective in sentence level sentiment classification [Zhang et al. 2010]. In addition to the classical machine learning approaches like NB and SVM, researchers also develop some other learning approaches for sentiment classification. A minimum-cut-based approach is employed to classify the movie reviews as subjective and objective in [Pang and Lee 2004]. In this work, the sentences are linked together with the association scores, which indicate how much degree the connected sentences will fall into one class. Then the minimum-cut algorithm is used to partition the sentence graph and to distinguish subjective movie reviews from those objective ones. It is concluded that the minimum-cut approach can improve the performance of subjective classification. [Mcdonald et al. 2007] apply the structured model, a model similar to conditional random fields (CRF) [Lafferty et al. 2001], to fine-to-coarse sentiment analysis. The proposed model is able to simultaneously identify the sentiment polarities of a document and its compositions including the sentences, phrases and words in the document. They apply the hidden conditional random field approach [Quattoni et al. 2007] to further pursue the fine-grained sentiment classification. [Bickerstaffe and Zukerman 2010] develop a decision tree of SVMs for document-level multi-class sentiment classification, which leverages the inter-class similarity in the learning process.

With the growth of the Web2.0 technologies, a large number of social media websites spring up on the Internet. This provides more opportunities for the public to express their opinions and feelings. To express strong opinions in social media, people often prefer to use special indicators such as emoticons ("☺", "☹", "^_^!"), punctuations ("*!!!!*", "*????*") and words with repeating letters ("*gooood*", "*baaad*"). Besides the previously mentioned textual features, e.g., the *N*-gram features, the features related to the writing style are also explored by the researchers for sentiment classification of social media text. [Melville et al. 2009; Gao et al. 2012b] focus on blogosphere sentiment classification and [Bakliwal et al. 2012; Aisopos et al. 2012; Davidov et al. 2010] focus on microblog sentiment classification.

2.1.1.3 Semi-supervised Sentiment Classification

Semi-supervised learning is another technique widely used in sentiment classification. Two problems have been recognized in supervised sentiment classification. First, the acquisition of the training data can be hard and expensive for a learning based approach. But it is often easy to get a large amount of unlabeled data. Second, sentiment classification is sensitive to the domain. For example, a word may be positive in one domain, but negative in another domain.

It needs to consider the domain adaption issue in multiple domain sentiment classification. To reduce the impact of these two problems in a certain degree, the semi-supervised approaches [Zhu et al. 2009; Chapelle et al. 2006] are widely adopted in cross-domain and cross-lingual sentiment classification [Brooke et al. 2009].

[Ju et al. 2012] focus on multiple domain sentiment classification. They propose a semi-supervised learning approach to minimize the annotation cost of data labeling. The proposed approach trains a classifier by actively selecting from amounts of unlabeled data those sentences that tend to be more valuable for classification, like the sentences containing the sentiment words. This approach can achieve a promising result by using only a few annotated data on multiple domain datasets. [Wan 2009] addresses the task of cross-lingual sentiment classification. Due to the lack of Chinese labeled data, he translates the labeled sentiment reviews from English to Chinese using public available machine translators. The co-training algorithm is then applied to both English and Chinese sentiment classification. The further developed approaches to cross-lingual sentiment classification are reported in [Lu et al. 2011; Meng et al. 2012a]. Another line of the cross-lingual sentiment classification research applies topic modeling approaches like Latent Dirichlet Allocation (LDA) [Blei et al. 2003] and its variants in classification. To model the sentiment polarity of each review, the LDA-like approaches assume that each review is a mixture of sentiment polarities and each polarity is a probability over the words. This idea is applied in fine-gain sentiment classification [Tackstrom and McDonald 2011], which aims to classify the sentiment polarity of a review as well as the sentiment polarities of the phrases and words in the review.

2.1.2 Sentiment Extraction

In practice, many applications require the systems to not only identify the sentiment polarity conveyed in the text unit, but also extract the specific sentiment segments from it. These extraction applications are generally recognized as *Information Extraction* (IE) in natural language processing. Here, opinion-oriented information extraction is referred to as *Sentiment Extraction*, which extracting the sentiment components, e.g., opinion holders, opinion targets, aspects of the opinion targets, and sentiment words as mentioned in Chapter 1. Most existing work on sentiment extraction focuses on the aspect extraction.

2.1.2.1 Rule-based Sentiment Extraction

The rule-based approaches extract the sentiment components according to handcrafted rules based on the syntactic patterns, the co-occurrences between the sentiment components and etc.

[Hu and Liu 2004] assume that the frequent nouns and noun phrases tend to express the product aspects. A POS tagger is employed to identify nouns and noun phrases in the product reviews. A threshold determined empirically is used to filter out the low-frequent nouns and noun phrases. This approach is further improved by [Popescu and Etzioni 2005]. They remove the noun phrases that may not related to the product aspects by computing the PMI measures between the phrases and certain *discriminator phrases*. The discriminator phrase is defined as the product category or the domain of the product. For instance, in the sentence "*XXXX* is a mobile phone", "mobile phone" is the category of the products. With the identified seed discriminator phrases, they derive the PMI measures of the noun phrases to the seed phrases. The noun phrases with low

PMI are not likely to be the sentiment components of the product review. [Moghaddam and Ester 2010] explore more sophisticated POS patterns. Based on POS patterns, they remove non-aspect nouns phrases that are not considered as the sentiment components. The similar work that extracts the sentiment components based on POS patterns is also reported in [Ku et al. 2006; Blair-Goldensohn 2008].

Intuitively, the sentiment components are concurrent in an opinionated text unit. Some existing approaches have adopted this idea in sentiment extraction. [Zhuang et al. 2006] use Stanford Parser[5] to identify the relations between the aspects and sentiment words in movie reviews. The keyword list and the relation templates are leveraged to extract the aspects and the sentiment words together. [Wu et al. 2009] define an opinion triple consisting of aspects, sentiment words and sentiment polarity, like {“iPhone camera”, “extremely good”, “positive”}. Instead of using the Stanford Parser, they propose a phrase-level dependency parsing approach to mine the triple (phrase level) dependency in the product reviews. Alternatively, [Qiu et al. 2011] identify sentiment words and the product aspects in a bootstrapping way. Another two bootstrapping approaches are also reported in [Kobayashi et al. 2006; Somasundaran and Wiebe 2009].

To sum up, these approaches can discover the explicit sentiment components, like the product aspects and the sentiment words that appear frequently in the reviews. The performance of these approaches highly depends on the coverage and the accuracy of the human-defined rules, and it is sensitive to the domains.

---

[5] http://nlp.stanford.edu/index.shtml

2.1.2.2 Learning-based Sentiment Extraction

Similar to information extraction [Mooney and Bunescu 2005; Sarawagi 2008; Hobbs and Riloff 2010], sentiment extraction can also be cast as a supervised learning problem. Given a text unit, learning-based sentiment extraction aims to predict the classes (e.g., opinion target, aspect or sentiment word) of the sentiment components in the text unit.

[Kobayashi et al. 2007] focus on sentiment aspect extraction. They select two kinds of features, the contextual clues (e.g., the *VP* pattern [6] <*feel comfortable*>) and the context-independent statistical clues (e.g., the statistical co-occurrence clues) from Web documents. Based on these features, they locate the candidates of the product aspects in the documents and then apply a tree-structured model to classify the candidates as aspects or not. [Yu et al. 2011] classify aspect candidates with one-class SVM [Manevitz and Yousef 2002]. [Kovelamudi and Ramalingam 2011] compute the semantic relevance of the two words by mapping them to Wikipedia [7] and then apply SVM to aspect classification.

Different from the above-mentioned classification approaches, some studies regard sentiment extraction as a sequential labeling problem and apply the sequential learning approaches like Hidden Markov Models (HMM) [Rabiner 1990; Ghahramani and Jordan 1997] and Conditional Random Fields (CRF) [Lafferty et al. 2001] for extraction. For example, [Jin and Ho 2009] integrate the syntactical features (e.g., POS) and the contextual features (e.g., the surrounding

[6] VP means verb phrase, which is a phrase syntactic category [Chomsky 2002].

[7] www.wikipedia.org

words of a given word). Then they propose a novel lexicalized HMM-based model to learn the sentiment components. Based on CRF, [Li et al. 2010a] propose a new machine learning approach, namely Skip-tree CRF, which not only considers the word sequential information, but also mines the sentence structural information. Similarly, the hierarchical CRF approach is studied in [Choi and Cardie 2010] to identify the opinion expressions and determine their sentiment polarities.

Among these learning-based sentiment extraction approaches, the sequential learning approaches are the dominant approaches and their effectiveness have been demonstrated in the previous work.

2.1.2.3 Topic Modeling based Sentiment Extraction

Topic model is a statistical model to uncover the hidden topics in the document collections [Blei 2012]. The topic modeling approaches, e.g., Probabilistic Latent Semantic Analysis (pLSA) [Hofmann 1999] and LDA [Blei et al. 2003; Griffiths and Steyvers 2003]) assume that the document is a mixture of topics and each topic is a distribution over words. Intuitively, the topics discovered by the topic models are most likely to be the opinion targets or the aspects of the opinion targets. Therefore, researchers explore the topic modeling based approaches for sentiment extraction, especially for sentiment aspect extraction.

[Mei et al. 2007] present a new probabilistic model, called Topic-sentiment mixture (TSM) model to simultaneously model the topical words and the sentiment words in blogospheres. While the TSM model is extended from pLSA, there are also some other studies extending the LDA model. [Titov and

McDonald 2008] address the problem of multi-grain sentiment analysis by combining the pLSA and LDA models. [Branavan et al. 2008] cluster the key-phrases like “a nice shoot” and “well done” into the semantic product features, and leverage the derived clusters to identify the hidden topics in the document collection.

In an opinionated text unit, the aspect words and the sentiment words usually appear concurrently. Topic modeling can discover the aspect words and the sentiment words together, but it cannot further tell which one is the aspect word and which one is the sentiment word. In [Brody and Elhadad 2010], the aspect words are identified by the topic model, and the sentiment words are identified by only taking the adjectives into consideration. [Zhao et al. 2010] propose a hybrid model which combines the Maximum Entropy (MaxEnt) model with an extended LDA model. The LDA topic model is to uncover the aspects words and sentiment words from the collection. The MaxEnt model then leverages the syntactic features to classify the aspect words and the sentiment words. [Sauper et al. 2011] propose a hybrid model by integrating the HMM and the LDA models. This hybrid model can identify the sentiment components in the reviews and identify the classes (aspect words or sentiment words) of the sentiment components simultaneously. A similar joint topic model is also proposed in [Lu and Zhai 2008; Jo and Oh 2011].

Currently, the joint model is widely studied in topic modeling based sentiment extraction. In the joint model, the topic modeling is used to discover the sentiment components, and the other model (e.g., MaxEnt, HMM) is used to classify the classes of the sentiment components identified from topic modeling.

2.1.3 Opinion Summarization

For the previous two research topics, the researchers mainly focus on sentiment analysis on a single review. However, some applications require exploiting the sentiment information embedded in the whole collection of sentiment-bearing texts. Opinion summarization is such a kind of research that automatically summarizes the sentiment information from a collection of sentiment-bearing texts and properly displays the generated summaries [Wang and Liu 2011]. Opinion summarization has been widely studied in industry like in the companies Microsoft[8], Google[9], Amazon[10] and K-matrix[11]. It can greatly benefit the business intelligence research.

Opinion summarization is a special case of document summarization which has been studied for about a half century [Das and Martins 2007; Nenkova and McKeown 2012; Gao et al. 2012a]. However, opinion summarization is different from traditional document summarization in three perspectives. (1) Unlike traditional document summarization, opinion summarization highly emphasizes the sentiment-oriented texts and needs to perform the syntactic and the semantic analysis on the selected sentences in order to extract the opinion targets, the aspects of the opinion targets and the explicit or implicit sentiments. (2) They both assume that the document(s) contain redundant information. The redundant information is excluded in the generated summary in traditional document summarization. Nonetheless, opinion summarization usually needs to include the

---

[8] www.microsoft.com
[9] www.google.com
[10] www.amazon.com
[11] www.kmatrixonline.com

quantitative information about the redundancy in the generated opinion summary. It is important to show how many people express the positive sentiments and how many people hold the negative sentiments. (3) By now, most studies on traditional document summarization generate summaries with the extractive techniques [Gupta and Lehal 2010]. However, opinion summarization can be regarded as an abstractive document summarization [Ganesan et al. 2010; Cheung 2008], which aims to generate structured summaries with the extracted opinion targets, the related aspects and the corresponding sentiment words.

2.1.3.1 Aspect-based Opinion Summarization

At present, aspect-based opinion summarization is the mainstream research in opinion summarization. It aims to summarize the aspects and the related sentiment polarities given the collection of reviews about a particular domain.

Feature class: OA
PRO: 70
Sentence 1: The movie is excellent.
Sentence 2: This is the best film I have ever seen.
...
CON: 10
Sentence 1: I think the film is very boring.
Sentence 2: There is nothing good with the movie.
...

Figure 2.1 Example of aspect-based opinion summarization from [Zhuang et al. 2006]

[Hu and Liu 2004] pioneer aspect-based opinion summarization on customer reviews. The generated summary likes the one presented in Figure 2.1, which displays the aspects, the corresponding sentiment polarities and the quantities of the text units which hold the sentiment polarities. [Ku et al. 2006] study opinion summarization together with opinion extraction and opinion tracking. They propose an algorithm to determine the relevance and sentiment scores of each extracted aspects. Based on these scores, they generate opinion summaries for news articles and blogospheres. [Nishikawa et al. 2010a] propose an algorithm that takes both the sentence content and the sentence coherence into consideration when generating the opinion summary. They further propose an Integer Linear Programming (ILP) approach to search for the optimal sentiment-oriented sentence sequence [Nishikawa et al. 2010b]. This sentence sequence is regarded as the opinion summary. Similar to graph-based document summarization [Mihalcea 2004; Otterbacher et al. 2005], a graph-based opinion summarization approach is proposed in [Ganesan et al. 2010], which determines the salience of each aspect based on an aspect graph.

Rather than generating the aspect-based opinion summary in Figure 2.1, [Kim and Zhai 2009] propose the contrastive opinion summary, which actually includes two sub-summaries with different sentiment polarities (i.e., positive and negative), respectively. [Paul et al. 2010] further purse the problem of contrastive opinion summarization. A bit difference is that their generated summary is a set of sentence pairs, and each sentence pair contains two sentences with different sentiment polarities.

Opinion summarization is still a hot research topic at present. It is also noteworthy that opinion summarization is not a self-governed topic in sentiment analysis. It is high-dependent on the researches in sentiment classification and sentiment extraction.

2.2 Sentiment Lexicon Learning

The studies on sentiment lexicon learning can be divided into *monolingual* *sentiment lexicon learning* and *cross-lingual sentiment lexicon learning*. Monolingual sentiment lexicon learning is the dominant research in sentiment lexicon learning. While the current monolingual sentiment lexicon learning work mainly focuses on English sentiment lexicon learning, sentiment lexicon learning for non-English languages is still under-researched.

2.2.1 Monolingual Sentiment Lexicon Learning

The objective of monolingual sentiment lexicon learning is to learn sentiment lexicons for only one language. There exist two main approaches in monolingual sentiment lexicon, the *co-occurrence based sentiment lexicon learning* approaches and the *semantic based sentiment lexicon learning* approaches.

2.2.1.1 Co-occurrence based Sentiment Lexicon Learning

The *co-occurrence based* approaches determine the sentiment polarity of a given word according to its statistical information e.g., the co-occurrence of the word to some pre-defined sentiment seed words.

One of the earliest work conducted by [Hatzivassiloglou and McKeown 1998] assumes that a conjunction word conveys the polarity relationship between the two words that it connects. For example, the conjunction word "*and*" tends to link the two words with the same polarity, while the conjunction word "*but*" is likely to link the two words with the opposite polarities. This work only considers adjectives, but not nouns or verbs. It can only deal with the situations where the two sentiment words are connected by specific conjunctions. But if a sentiment word always appears in the dataset alone, the approach will fail to extract this sentiment word. The same approach is used in [Xu et al. 2010] but for Chinese sentiment lexicon learning. [Riloff et al. 2003] define several pattern templates to extract sentiment words using bootstrapping approaches. [Turney and Littman 2003] calculate the PMI of a given word with the positive and negative sets of sentiment words. The sentiment polarity of the word is determined by the average PMI values to the positive and negative sets. To obtain PMI, they provide queries consisting of the given word and the sentiment word to the search engine. The number of hits and the positions (if the given word is near to the sentiment word) are used to estimate the association of the given word to the sentiment word. [Hu and Liu 2004] study sentiment word learning on customer reviews. They assume that the sentiment words tend to be correlated with the product features. Treating the frequent nouns and the noun phrases as the product features, they extract the adjective words as sentiment words from those sentences that contain one or more product features. This approach may work on the product reviews, where a product feature may frequently appear. But for other documents like news articles, this approach may not be effective. [Qiu et al. 2011] combine sentiment lexicon learning and opinion target extraction. A bootstrapping approach is proposed to learn the sentiment words and to extract the opinion targets simultaneously based on eight manually-defined rules.

2.2.1.2 Semantic based Sentiment Lexicon Learning

The *semantic based* approaches determine the sentiment polarity of a given word according to its semantic relations (e.g., the synonym relations) to the sentiment seed words. The word semantic relations can be obtained from the public dictionaries like *WordNet*.

[Kim and Hovy 2004] assume that the synonyms of a positive (or negative) word are also positive (or negative) but the antonyms of it are negative (or positive). Initializing with a set of sentiment words, they expand the existing sentiment lexicons based on these two kinds of word relations. [Kamps et al. 2004] build a synonym graph according to the synonym relations (i.e., synset) derived from WordNet. The sentiment polarity of a word is then calculated based on the shortest path from it to the two sentiment words "*good*" and "*bad*". However the shortest path may not precisely describe the sentiment orientation because it has been noticed that there are only five steps between the word "*good*" and the word "*bad*" in WordNet [Hassan et al. 2011]. [Takamura et al. 2005] construct a word graph based on the WordNet glosses (i.e., the textual definitions of the words in WordNet). The two words are connected if one word appears in the glosses of the other. A spin model is proposed to determine the word sentiment polarity with the word graph. [Rao and Ravichandran 2009] build a word graph through the word relations (e.g., synonym, hypernym) in WordNet. Based on the word graph, they propose three graph-based semi-supervised approaches, e.g., minimal-cuts, randomized minimal-cuts and label propagation, to identify the sentiment polarity of the given word. [Esuli and Sebastiani 2006; Esuli and Sebastiani 2007; Baccianella et al. 2010] treat sentiment word learning as the machine learning problem and classify the sentiment orientations of the words in WordNet. They select seven positive words and seven negative words, and expand them through the "*see-also*" and "*antonym"* relations in WordNet. These expanded words are then used to train a ternary classifier to predict the sentiment polarities of all the words in WordNet. The features used for classification are extracted from the WordNet glosses. The generated sentiment lexicon is the well-known sentiment lexicon, *SentiWordNet*[12].

2.2.2 Cross-lingual Sentiment Lexicon Learning

Cross-lingual sentiment lexicon learning aims to learn the sentiment lexicon for one language based on the sentiment resources available in the other language(s). Usually, English is regarded as the source language as there are rich English sentiment resources, and the language for which we would like to build the sentiment is regarded as the target language.

[Mihalcea et al. 2007] generate the sentiment lexicon for Romanian by directly translating the English sentiment words into the corresponding Romanian words based on the bilingual English-Romanian dictionaries. When confronting multi-word translations, they validate the multiple word-to-word translations by the AltaVisa[13] search engine. A valid multiple word-to-word translation must occur at least three times on the Web. The approach proposed by [Hassan et al. 2011] learns the sentiment words based on English WordNet and WordNets in the target languages, i.e., Hindi or Arabic. The cross-lingual dictionaries are used to connect the words in the two languages. The polarity of a given word is determined by the average hitting time from the word to the English sentiment words. What these two approaches have in common is that they both connect the words in the two languages based on the cross-lingual dictionaries. The main concern of these approaches is the effect of morphological inflection, i.e., a word may be mapped to multiple words in the cross-lingual

---

[12] http://sentiwordnet.isti.cnr.it
[13] http://search.yahoo.com/?fr=altavista

dictionaries. For example, one single English word typically has four Spanish or Italian word forms (two each for gender and for number) and many Russian word forms (due to gender, number and case distinctions) [Steinberger et al. 2011]. Usually it needs an additional process to disambiguate the sentiment polarities of all the morphological variants.

To improve the sentiment classification performance for the target language, [Banea et al. 2010] translate the English sentiment lexicon to the target language with *Google Translator*[14]. Similarly, Google Translator is used by [Steinberger et al. 2011], who manually produce high-level gold-standard sentiment lexicons for two languages (e.g. English and Spanish) and then translate them into a third language (e.g., Italian) using Google Translator. They believe that the words in the third language, which appear in the both translation lists, are more likely to be the sentiment words. Both [Banea et al. 2010] and [Steinberger et al. 2011] connect the words in the two languages based on machine translation (MT) engines. The bottleneck of these approaches is low overlap between the vocabularies in human-written documents and the vocabularies in the documents translated by MT engines [Duh et al. 2011; Meng et al. 2012b]. It will lead to the low coverage.

The work on cross-lingual sentiment lexicon learning is still at the early stage. This dissertation focuses on cross-lingual sentiment lexicon learning. We plan to automatically learn sentiment lexicons for given languages by leveraging the existing sentiment resources (e.g., sentiment lexicons in English) from another languages. Meanwhile, we propose the generic approaches which are

[14] http://translate.google.com

language-independent and can learn the sentiment lexicons for any given languages.

2.2.3 Relation to the Other Research in Sentiment Analysis

By now, I have introduced the main lines of research in sentiment analysis. The following subsection will present the relationship between sentiment lexicon learning and the other sentiment analysis researches.

2.2.3.1 Relation to Sentiment Classification

Sentiment lexicon learning can be regarded as sentiment classification at the word level. Its relationships to document/sentence level sentiment classification are elaborated below.

Sentiment lexicon learning is a greater benefit to document/sentence level sentiment classification. For unsupervised sentiment classification, the word sentiment is the prerequisite. Some unsupervised approaches in sentiment classification regard sentiment lexicon learning as a subtask; some just use the sentiment words in existing sentiment lexicons to determine the sentiment polarity of the sentence or the document. For (semi-)supervised sentiment classification, the sentiment words can be directly used as representative features [Zhang et al. 2010].

Furthermore, sentiment lexicon learning is also different from document/sentence level sentiment classification in two perspectives. First, the text unit to be classified is different. Not surprisingly, the text unit in sentiment classification is a document or a sentence, while the text unit in sentiment lexicon learning is a word or a phrase. Second, for a text unit in sentiment classification, its sentiment polarity is unique, at least from the perspective of the writer. For a word, its sentiment polarity is related to the opinion target as well as to the aspects of the opinion target. Though a word can have different sentiment polarities, current work on sentiment lexicon learning usually assumes that there is an overall sentiment polarity for the word. This overall sentiment polarity of the word is shown in the learned lexicon.

2.2.3.2 Relation to Sentiment Extraction

Sentiment lexicon learning can be also regarded as a subtask of sentiment extraction, which means to extract the sentiment words from the corpora. In the past, researchers have already proposed the approaches to extracting sentiment words and aspect words simultaneously. For example, [Qiu et al. 2009; Qiu et al. 2011] propose a bootstrapping approach to generate the sentiment words and the product aspects iteratively. Thus, the conclusions can be safely drawn that sentiment extraction and sentiment lexicon learning are two highly related tasks and that the improvement in one task can help to improve the other.

2.2.3.3 Relation to Opinion Summarization

Opinion summarization is not an independent task and the performance of the generated opinion summaries is related to all the other tasks in sentiment analysis including sentiment classification and sentiment extraction. Combining the conclusions from the previous two sections, we believe that opinion summarization is implicitly related to sentiment lexicon learning as well.

2.2.4 Available Sentiment Lexicon

By now, there are several publically available sentiment lexicons.

- Harvard General Inquirer (GI) lexicon[15] [Stone et al. 1963]
- Micro-WNOp[16], a subset of WordNet [Cerini et al. 2007]
- Bing Liu's Sentiment lexicon[17] [Hu and Liu 2004]
- MPQA subjectivity lexicon[18] [Wilson et al. 2005]
- SentiWordNet [Esuli and Sebastiani 2006; Esuli and Sebastiani 2007; Baccianella et al. 2010]
- HowNet[19] (Chinese) [Dong et al. 2010]
- Linguistic Inquiry and Word Counts (LIWC)[20]

Among these sentiment lexicons, GI and Micro-WNOp are human-created lexicons.

2.3 Chapter Summary

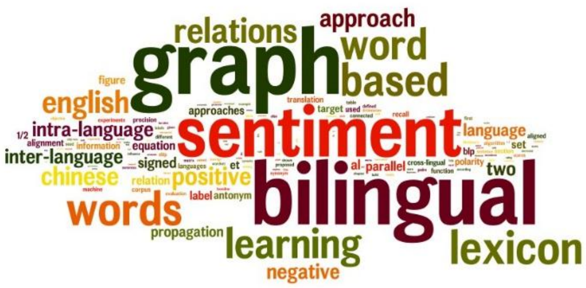
I briefly introduce the main researches in sentiment analysis in this chapter. I present the state-of-the-art studies on sentiment classification, sentiment extraction and opinion summarization and described the main techniques explored in these studies. I then review the related work in sentiment lexicon learning and analyze the relationships between sentiment lexicon learning and the other related work in sentiment analysis. From the literature review and discussion, it is clear that sentiment lexicon learning plays an important role in

---

15 www.wjh.harvard.edu/~inquirer
16 www-3.unipv.it/wnop
17 www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
18 www.mpqa.cs.pitt.edu
19 www.keenage.com
20 www.liwc.net

sentiment analysis.

Chapter 3 Bilingual Graph based Sentiment Lexicon Learning



3.1 Chapter Overview

In this chapter, my first work on cross-lingual sentiment lexicon learning is presented. I name this work as *bilingual graph based sentiment lexicon learning*, which is defined to automatically generate the sentiment lexicon for a given language based on a bilingual graph. The bilingual graph consists of the words in the given language (aka. the target language) and the words in another language where there exist rich sentiment resources (aka. the source language). Considering that currently English is rich in available sentiment resources, English is included as the source language. Then the objective of bilingual graph based sentiment lexicon learning is to transfer the sentiment information from English to the target language.

To transfer the sentiment information between languages, it is required to set up the relations between the English words and the words in the target language. Then the sentiment polarities of the words in the target language are determined according to the word relations and the English sentiment seed words. In these two steps, the relation construction plays a fundamental role in the learning processes since the relations between the words in the two languages are responsible for the transfer of sentiment information between the languages. Two methods are often used to connect the words in different languages in the literature. One is based on translation entries in bilingual dictionaries [Hassan et al. 2011]. The other relies on a machine translation (MT) engine as a black-box to translate the sentiment words in English to the target language [Steinberger et al. 2011]. However, as reported in [Mihalcea et al. 2007; Duh et al. 2011], these approaches tend to use a small set of vocabularies to translate or to map the natural languages, which leads to a low recall of the generated sentiment lexicons for the target language.

To solve this problem, I propose a generic approach to address bilingual graph based sentiment lexicon learning. Specifically, I formalize this task as bilingual graph modeling, in which the intra-language relations among the words in the same language and the inter-language relations among the words between different languages are properly represented. The intra-language relations are used to model the semantic relations (such as synonym and antonym) among the words in the same language, while the inter-language relations are used to bridge the words from different languages and transfer the sentiment information from English to the target language. Two types of word graphs, i.e., the bilingual graph and the signed bilingual graph are explored. The difference between them is that in the bilingual graph the weights of the relations (including the intra-language and inter-language relations) are all positive, while in the signed bilingual graph the weights of the relations can be either positive or negative. Based on these two types of graphs, two label propagation-based approaches are developed to induce the sentiment lexicon for the target language by using the words in English sentiment lexicon as seeds. The propagation algorithms simultaneously take the intra-language relations and the inter-language relations into account in an iterative way, which is proved to greatly improve the precision of the results. Moreover, I propose to use the word alignment information derived from a parallel corpus to construct the inter-language relations in the (signed) bilingual graph. The two words from different languages that are aligned to each other in a parallel sentence pair are connected. Taking advantage of a huge amount of parallel corpus, this approach significantly improves the coverage of the generated sentiment lexicon.

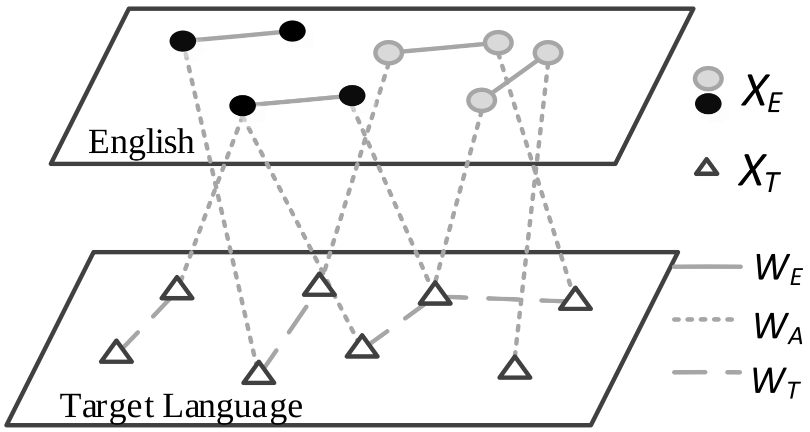


Figure 3.1 Bilingual graph for CSLL

As will be explicated in Section 3.2, the bilingual graph is established with the words in the source and target languages and with the well-represented intra-language and inter-language relations. A bilingual graph label propagation approach is presented to learn the sentiment lexicon for the target language based on the bilingual graph. In Section 3.3, a signed bilingual graph is built by further incorporating the antonym intra-language relations into the bilingual graph. The approach introduced in Section3.2 is further extended for learning sentiment lexicons based on the signed bilingual graph. Finally, Section 3.4 presents the summary of this chapter.

3.2 Bilingual Graph based Sentiment Lexicon Learning

3.2.1 Formalization of Bilingual Graph based Sentiment Lexicon Learning

Given the words in English and in the target language, we can build a bilingual graph as illustrated in Figure 3.1. Based on such a graph, the purpose of bilingual graph based sentiment lexicon learning is to induce the sentiment labels of the words in the target language based on the English sentiment seed words, represented by the shaded nodes in the English sub-graph in Figure 3.1. Here, this task is approached using a Bilingual graph Label Propagation (**BLP)** approach.

Specifically, the bilingual graph is represented by $G = (X_E \cup X_T, W_E \cup W_T \cup W_A)$, which consists of the English and Chinese sub-graphs $G_E = (X_E, W_E)$ and $G_T = (X_T, W_T)$, where X_E and X_T denote the labeled words in English and the unlabeled words in the target language, and W_E and W_T represent the intra-language relations of the words in English and the intra-language relations of the words in the target language. These two intra-language sub-graphs are connected by a bipartite graph $G_A = (X_E \cup X_T, W_A)$, where W_A represents the inter-language relations between the words in English and the target language. In this bilingual graph, all the relations have positive values, that is $W_A, W_E, W_T \in \mathbb{R}_+$. With labels Y_E for the sentiment seed words X_E, the BLP approach predicts the labels of the words X_T. The details of the bilingual graph construction process and the bilingual graph label propagation algorithm are introduced in the following sections.

3.2.1.1 Bilingual Graph Building

The words in both English and the target language are represented by the nodes in the graph. The synonym relations of the words in the same language are leveraged to build the intra-language relations.

To build the inter-language relation, there are two intuitive ways to connect the words in the two languages. One way is to insert a links to the two words if there exists an entry mapping between the words in a bilingual dictionary (e.g., the Universal Dictionary[21] English-Chinese dictionary) (see Figure 3.2(b)). This method is simple and straightforward, but has two limitations. 1) Dictionaries are static in a certain period, while the sentiment words evolve over time frequently. 2) The entries in dictionaries are normally the expressions for the formal and written languages, but people prefer using the colloquial language in expressing their sentiments or opinions. These limitations lead to the low coverage of the links from English to the target language [Steinberger et al. 2011]. An alternative way is to use a machine translation engine as a black-box to build the inter-language relations (see Figure 3.2(c)). One can send a word in English to a public available machine translation engine and get the translations in the target language [Mihalcea et al. s2007]. Edges are then inserted into the graph between

---

[21] www.dicts.info/uddl.php

the English word and its corresponding translations. This approach suffers from the problem of the low coverage as well since machine translation engines tend to use a small set of vocabularies to translate the natural languages [Duh et al. 2011; Meng et al. 2012].

Figure 3.2 (a) Parallel sentence; (b) Bilingual dictionary; (c) Machine translation

In this work, I propose to leverage a large bilingual parallel corpus, which is readily available in the machine translation research community, to set up the bilingual graph. The parallel corpus consists of millions of parallel sentence pairs from two different languages, which have been used as the foundation of the state-of-the-art statistical machine translation engines. In the example shown in Figure 3.2(a), the two sentences in English and Chinese are parallel sentences, which express the same meaning in different languages. It is easy to derive the word alignment from the sentence pairs automatically using a tool that implements with a state-of-the-art approach, like GIZA++ [22] or BerkeleyAligner[23]. In this example, the Chinese word "快乐" (happy) is linked to the English word "happy" and we say that these two words are aligned. Similarly the English words "best" and "wishes" are both aligned to "祝" (wish).

The word alignment information encodes the rich association information between the words from the two languages. I am therefore motivated to leverage the parallel corpus and word alignment to construct the bilingual graph for cross-lingual sentiment lexicon learning. The words from both languages in the bilingual parallel corpus are taken as the nodes in the bilingual graph. The inter-language relations are built by connecting the two words that are aligned together in a parallel sentence pair. There are several advantages of using the parallel corpus to establish the bilingual graph. First, the large parallel corpora have been extensively used in training the statistical machine translation engines and are easily reused for our task. The parallel sentence pairs are usually automatically collected and mined from the Web. As a result, they contain the different and practical variations of the words and phrases embedded in the sentiment expressions. Second, the parallel corpus is dynamically changed when necessary since it is relatively easy to collect from the Web. As a result, the novel sentiment information inferred from the parallel corpus is able to update the existing sentiment lexicons easily. These advantages greatly improve the recall of the generated sentiment lexicon, as demonstrated later in our experiments.

---

[22] www.statmt.org/moses/giza/GIZA++.html
[23] http://nlp.cs.berkeley.edu

3.2.1.2 Bilingual Graph Label Propagation

The early work on label propagation is usually based on a single graph [Zhu and Ghahramani 2002; Zhou et al. 2003]. Zha et al. incorporate the class graph into the learning processing and develop a multi-label label propagation approach [Zha et al. 2009]. Later on, researchers develop some similar approaches and apply them to the applications, like Part-Of-Speech tagging [Das and Petrov 2011; Li et al. 2012], image annotation [Wang et al. 2011] and protein function prediction [Jiang 2011; Jiang and McQuay 2012].

In this section, I present a novel bilingual graph label propagation approach for cross-lingual sentiment lexicon learning. Let F denote the prediction function, which generates the labels Y for the unlabeled words X. The objective function of the BLP approach is defined as:

$$argmin(\Omega(F)) = argmin(\Omega_l(F) + \Omega_s(F)) \quad \text{E3.1}$$

where the *loss* function $\Omega_l(F)$ means that the prediction should not change too much from the initialized label assignment. The *smoothness* function $\Omega_s(F)$ requires the nearby nodes to share the same labels.

In the bilingual graph, the loss function $\Omega_l(F)$ is further defined as:

$$\Omega_l(F) = \mu \sum_{i=1}^{n} \left\| f_{E_i} - y_{E_i} \right\| \quad \text{E3.2}$$

The smoothness function $\Omega_s(F)$ is defined by the sum of the intra-language smoothness $\Omega_s^{intra}(F)$ and the inter-language smoothness $\Omega_s^{inter}(F)$, which are further defined as

$$\Omega_s^{intra}(F) = \frac{1}{2}\rho_1 \sum_{i,j=1}^{n} w_{E_{ij}} \left\| \frac{f_{E_i}}{\sqrt{d_{E_{ii}}}} - \frac{f_{E_j}}{\sqrt{d_{E_{jj}}}} \right\| + \frac{1}{2}\rho_2 \sum_{i,j=1}^{m} w_{T_{ij}} \left\| \frac{f_{T_i}}{\sqrt{d_{T_{ii}}}} - \frac{f_{T_j}}{\sqrt{d_{T_{jj}}}} \right\| \quad \text{E3.3}$$

and

$$\Omega_S^{inter}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{A_{ij}} \left\| \frac{f_{E_i}}{\sqrt{d_{AL_{ii}}}} - \frac{f_{T_j}}{\sqrt{d_{AR_{jj}}}} \right\|$$

D_E and D_T are two degree matrices of the intra-language relations W_E and W_T, respectively. They are the diagonal matrices, of which the (p,p) element $d_{E_{pp}}$ (or $d_{T_{pp}}$) is the sum of the p row of W_E (or W_T). The degree matrices D_{AL} and D_{AR} of the inter-language relations W_A are calculated by

$$D_{AL} = diag(\sum_j w_{1,j}, \sum_j w_{2,j}, \dots, \sum_j w_{n,j})^{\mathrm{T}}$$
$$D_{AR} = diag(\sum_i w_{i,1}, \sum_i w_{i,2}, \dots, \sum_i w_{i,m})^{\mathrm{T}} \qquad \text{E3.4}$$

To obtain the solution of Equation E3.1, the objective function is differentiated and then we have

$$\left.\frac{\partial(\Omega(F))}{\partial(F)}\right|_{F=F^*} = F^* - SF^* + \mu(F^* - Y) = 0 \qquad \text{E3.5}$$

where F^* is the optimal solution of Equation E3.1. The affinity matrix S is the combination of S_E, S_T and S_A, shown in Equation E3.6.

$$\mathrm{S} = \begin{bmatrix} (1-\beta)S_E & \beta S_A \\ \beta S_A^{\mathrm{T}} & (1-\beta)S_T \end{bmatrix} \qquad \text{E3.6}$$

The affinity matrixes S_E (for English) and S_T (for target language) are equal to $D_E^{-\frac{1}{2}} W_E D_E^{-\frac{1}{2}}$ and $D_T^{-\frac{1}{2}} W_T D_T^{-\frac{1}{2}}$. The inter-language relation S_A is equal to $D_{AL}^{-\frac{1}{2}} W_A D_{AR}^{-\frac{1}{2}}$.

The user-defined parameter $\beta \in (0,1)$ is used to adjust the relative importance of intra-language and inter-language propagation. By Equation E3.5, the optimization solution is obtained, i.e.,

$$F^* = \alpha(I - \gamma S)^{-1} Y \qquad \text{E3.7}$$

where $\alpha = \frac{1}{(1+\mu)}$ and $\gamma = \frac{\mu}{(1+\mu)}$.

Algorithm 3.1 Bilingual graph label propagation

Input: Given $G = (X_E \cup X_T, W_E \cup W_T \cup W_A)$. label Y_E for X_E, initialize μ and β

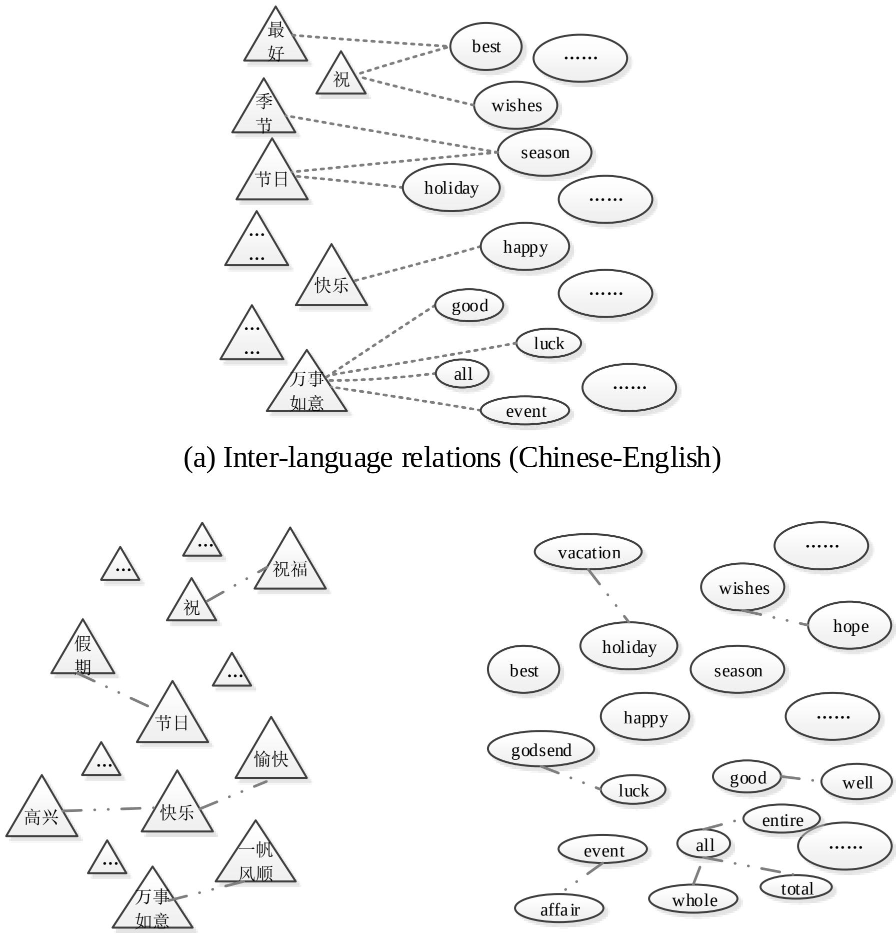
Output: Label F_T for X_T

1. Initialize Y_E with the English sentiment seeds
2. Set Y_T as zero and generate Y_0 with Y_E
3. Construct S according to Equation E3.6
4. Loop
5. $\quad F(t+1) = (1-\delta)SF(t) + \delta Y_0$
6. Until Y converge

To avoid the computation of the inverse matrix in Equation E3.7, the iteration algorithm is adopted in Algorithm 3.1. In Line 1, the label of the positive seed x_i is set to y_{E_i}= (1, 0) and the label of the negative seed x_j is set to y_{E_j}= (0, 1). Line 2 sets Y_T as a zero matrix and then generates Y_0 by combining Y_E with Y_T. In Line 3, the affinity matrixes S_E, S_T and S_A are combined together according to Equation E3.6. The positive and negative information is simultaneously propagated through Lines 4 to 6 until the predicted labels Y are converged. It has been proved that the iterative algorithm can converge to Equation E3.7 finally [Zhou et al. 2003].

In summary, the bilingual graph $G = (X_E \cup X_T, W_E \cup W_T \cup W_A)$ is constructed based on the words X_E and X_T, which are connected by the word alignments W_A derived from the parallel sentences (see Figure 3.3(a)). Within the same language, the words are connected with the synonym relations, W_E

and W_T, in WordNet (see Figure 3.3(b) and Figure 3.3(c)). The BLP approach generates the sentiment labels F (either positive or negative) for the unlabeled words in X_T. More precisely, for the word x_i, if $|f(i,0) - f(i,1)| < \xi$ (ξ is set to 1.0E-5), x_i is regarded as neutral. If $(f(i,0) - f(i,1)) \geq \xi$, x_i is reckoned to a positive word. And if $(f(i,1) - f(i,0)) \geq \xi$, x_i is assessed to be negative.



(a) Inter-language relations (Chinese-English)

(b) Intra-language relations (Chinese)

(c) Intra-language relations (English)

Figure 3.3 Bilingual graph building

3.2.2 Experiments

The experiments on Chinese sentiment lexicon learning are conducted to verify the proposed approach. Similar to the previous work [Rao and Ravichandran 2009; Baccianella et al. 2010], the sentiment words in the GI lexicon [Stone et al. 1963] are selected as the English seeds. In total, 2,005 positive words and 1,635 negative words are collected from the GI lexicon. To construct the bilingual graph, the ISI Chinese-English parallel corpus[24] is used, which contains the news articles published by *Xinhua News Agency* in Chinese and English collections [Munteanu and Marcu 2005]. Altogether, more than 25 million parallel sentence pairs in English and Chinese are collected. The stop-words in Chinese and English (e.g., "的" (of) and "am") together with the low-frequency words whose occurrence times are less than 5 are removed during preprocessing. Finally we get more than 174K English words (among which 3,519 words are with the sentiment labels) and more than 146K Chinese words (for which we need to predict the sentiment labels). The unsupervised method, namely *BerkeleyAligner*, is employed to align the parallel sentences in the ISI parallel corpus [Liang et al. 2006]. As an unsupervised method, it does not require to manually collect training data and does not need the complex training processing, yet the performance is still competitive with the supervised methods. Considering these two advantages, the *BerkeleyAligner* aligner is employed to generate the word alignments in the experiments. The word alignment frequencies are used to initialize the weights of the inter-language relations. The English and Chinese versions of WordNet[25] are employed to build the intra-language relations W_E and W_T, respectively. WordNets [Miller 1995] group the words into the synonym sets, called the synsets. Totally 117K synsets

[24] http://catalog.ldc.upenn.edu/LDC2007T09
[25] http://www.globalwordnet.org/gwa/wordnet_table.html

and 80K synsets are collected from English and Chinese WordNet, respectively.

The proposed approach first generates both positive and negative scores for each unlabeled word and then determines the word as positive or negative based on these two scores. The two sets of the newly-labeled positive and negative sentiment words are then ordered according to their polarity scores. The top ranked Chinese words are shown in Table 3.1.

Word	Meaning	Polarity	Word	Meaning	Polarity
好	Good	Positive	灾难	Disaster	Negative
正确	Correct	Positive	悲剧	Tragedy	Negative
有用	Useful	Positive	危险	Dangerous	Negative
聪明	Smart	Positive	伤害	Harm	Negative
高兴	Happy	Positive	错误	Fault	Negative
可靠	Reliable	Positive	愤怒	Rage	Negative
准确	Accurate	Positive	失败	Fail	Negative
有效	Effective	Positive	破坏	Damage	Negative
可爱	Cute	Positive	孤独	Alone	Negative
快乐	Happy	Positive	冲突	Clash	Negative

Table 3.1 Top ranked Chinese sentiment words

Precision@K is used to evaluate the ranked sentiment word lists. Two human annotators are also invited to annotate the top 1000 words produced by each approach. For P@10K, we separate the top 10K ranked list into ten equal parts sequentially. 100 words are randomly selected from each part. The human annotators are then asked to annotate all the selected words. The precision of these selected words are regarded as P@10K. Similar to the evaluation in TREC

Blog Distillation [Ounis et al. 2010], the human-annotated words from all the approaches are regarded as the benchmark sentiment words. Then based on the benchmark data, the ranked lists are evaluated in term of recall.

3.2.2.1 Comparison with Baseline Approaches

The following experiments are conducted to compare the proposed approach with the various baseline and existing approaches.

RULE: This baseline approach adopts the idea from Hu and Liu [Hu and Liu 2004], which assumes the synonyms of a positive (or negative) word are also positive (or negative). It regards the Chinese word that is aligned to the positive (or negative) English words as positive (or negative). If a word connects to both positive and negative English words, it is determined to be objective. Two sets of sentiment words are generated based on this simple heuristic.

SOP: This approach predicts the sentiment polarities of the unlabeled words based on the mean hitting time to the two sets of sentiment seed words [Hassan et al. 2011]. Given the graph $G = (X_E \cup X_T, W_E \cup W_T \cup W_A)$, it defines the transition probability from the word i to the word j as

$$p(j|i) = \frac{w_{i,j}}{\sum_k w_{i,k}} \quad \text{E3.8}$$

The mean hitting time $h(i|j)$ is the average number of the weighted steps from the word i to the word j. Starting from the word i and ending at a sentiment word $k \in \mathrm{M}$, the mean hitting time $h(i|M)$ is formally defined as

$$h(i|\mathrm{M}) = \begin{cases} 0, & i \in M \\ \sum_{j \in V} p(j|i)h(j|M) + 1, & otherwise \end{cases} \quad \text{E3.9}$$

Let M_+ and M_- denote the GI positive and negative seeds. If $h(i|M_+)$ is greater than $h(i|M_{-})$, the word x_i is classified as negative. Otherwise it is classified as positive. The generated positive words and negative words are then ranked according to their polarity scores, respectively.

Chinese	Positive	Negative	Average
RULE	0.360	0.363	0.362
SOP	0.516	0.486	0.501
BLP	0.702	0.706	0.704

Table 3.2 Comparison with baseline approaches (recall)

The recalls of these approaches are shown in the above Table 3.2. First, the recall of the RULE approach is significant lower than the recalls of the BLP and SOP approaches. Since many words in the corpus are aligned to both positive and negative words, it is not strange to have the low coverage with the RULE approach. For example, in most cases the positive Chinese word “帮助” (helpful) is aligned to the positive English word “helpful”. But sometimes it is also aligned (or misaligned) to the negative English words, like “freak”. Based on the heuristic, the word is predicted as objective. As a result, the RULE approach learn much fewer sentiment words compared with the other two approaches. With the SOP approach, the positive and negative scores are calculated according to the shortest paths between the unlabeled words and the positive or negative seed words. However, the shortest paths is usually coarse-grained to depict the sentiment polarity. You will find that there are only five steps between the word “*good*” and the word “*bad*” in WordNet.

The learned Chinese polarity word lists are then evaluated by precision at *k*. As illustrated in Figure 3.4, the BLP approach significantly outperforms the two baselines. The p values against the RULE and SOP approaches are 1.09E-8 and 1.4E-4, which indicates that the proposed BLP approach achieves a significant improvement. These results further indicate that the BLP approach is more effective in cross-lingual sentiment lexicon learning.

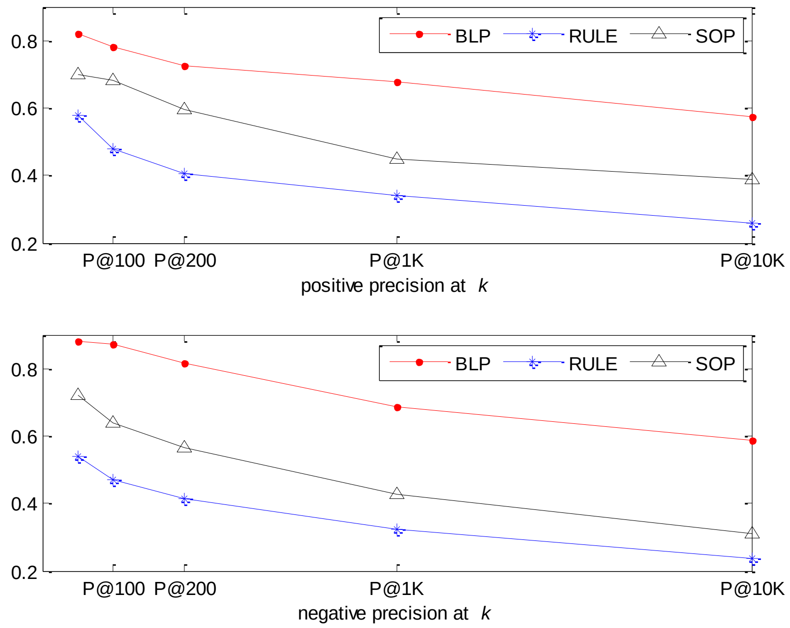


Figure 3.4 Comparison with the baseline approaches (precision)

The major differences between the BLP approach and the baselines are two-fold. First, in the baseline approaches the polarity information mainly transfers from English to Chinese and once a word gets a polarity score, the score is never changed or refined. In the BLP approach, the polarity information is transferred from English to Chinese and from Chinese to Chinese at the same time. Second, the polarity score of a word is updated with its connected words at each iteration. Thus the polarity information interacts more extensively and precisely in the BLP approach. To summarize, two aspects influence the learning process. 1) The inter-language relations influence the transfer of sentiment information between English and Chinese. 2) The intra-language relations have influence on sentiment propagation within each language.

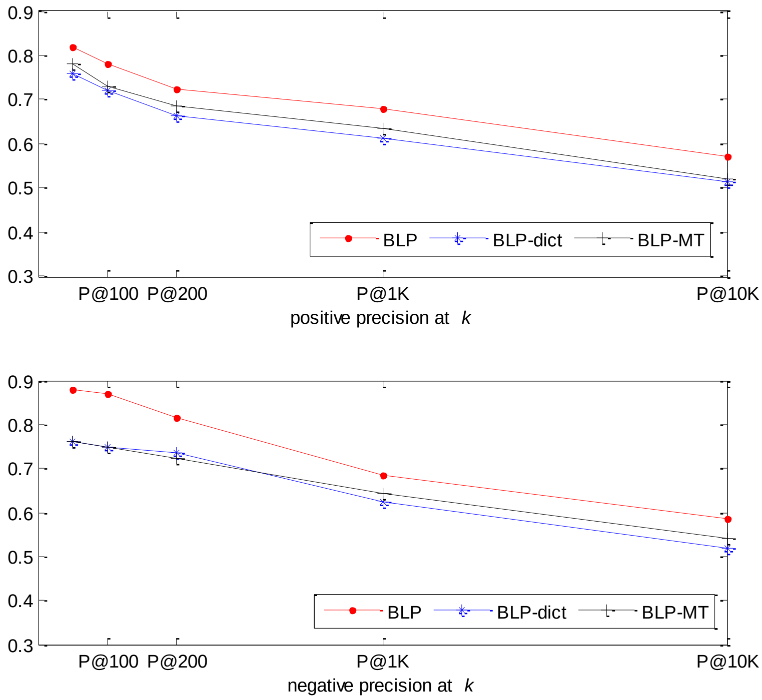


Figure 3.5 Evaluation on the inter-language relations (precision)

3.2.2.2 Evaluation on Inter-language Relation

This set of experiments aims to examine the selection of the inter-language relations.

BLP-dict: The inter-language relations are built upon the translation entries from LDC [26] and Universal Dictionary (UD). These dictionaries (both English-Chinese and Chinese-English dictionaries) contain more than 41K translation entries between the English and Chinese words. If the English word

[26] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm

x_i is translated to the Chinese word x_j in the dictionaries, $w_A(i,j)$ and $w_A(j,i)$ are set to ones.

BLP-MT: All the Chinese (English) words are translated into English (Chinese) by Google Translator. If the Chinese word x_i is translated to the English word x_j, the $w_A(i,j)$ and $w_A(j,i)$ are set to ones. If a Chinese word is translated to an English phrase, we assume that the Chinese word is projected to every word in the English phrase.

The learned Chinese sentiment word lists are also evaluated according to precision at *k*. Two findings are observed from the results shown in Figure 3.5. First, the BLP-dict and BLP-MT approaches outperform the two baseline approaches described in Section 3.2.2.1, which reconfirms the effectiveness of the proposed BLP approach in cross-lingual sentiment lexicon learning. Second, the alignment-based approach outperforms the dictionary-based and MT-based approaches. The reason contributing to this finding may be that more inter-language relations are brought in by the word alignments, compared with the translation entries from the dictionary and the translation pairs from the machine translator. For example, the English word "move" is often translated to "移动" (shift) and "感动" (affect, touch) by dictionaries or machine translation engines. From the parallel sentences, besides these word translation pairs, the word "move" is also aligned to "一帆风顺" (plain sailing bon voyage) that is commonly used in Chinese greeting texts. This translation entry is difficult to be discovered in dictionaries or by machine translation engines. In parallel sentences the words are aligned in the context of the sentence pairs. Sometimes the word "move" may be forced to be aligned to "一帆风顺" in the parallel sentences "good luck and best wishes on your career move" and "祝 | 你 | 新

的 | 事业 | 一帆风顺”. Thus when building the inter-language relations with word alignments, the BLP approach is likely to learn more sentiment word candidates. This also explains why the recalls of the dictionary-based and MT-based approaches are lower than that of the proposed BLP approach, as compared in Table 3.3. According to the statistic, in average a Chinese word connects to 2.3 or 1.7 English words if the inter-language relations are created with Google Translator or the dictionaries. But building the inter-language relations with word alignments, a Chinese word connects to 16.21 English words, which greatly increases the recall of our approach.

Chinese	Positive	Negative	Average
BLP-dict	0.647	0.654	0.651
BLP-MT	0.652	0.659	0.656

Table 3.3 Evaluation on the inter-language relations (recall)

To conclude, building the inter-language relations with word alignments improves the performance of cross-lingual sentiment lexicon learning. The word alignment based inter-language relations are able to largely convey the sentiment information between languages and able to provide much more sentiment word candidates.

3.2.2.3 Evaluation on Intra-language Relation

The next set of experiments is to examine the influence of the intra-language relations.

BLP-A: As the baseline of this set of experiments, it does not build the intra-language relations with either English or Chinese WordNet synsets. Only the inter-language relations with word alignments are used to construct the graph. That means to define W_E and W_T as zero matrixes.

BLP-AE: Different from BLP-A, besides word alignments, English WordNet synsets are used to build the intra-English relations W_E, but the intra-Chinese relations W_T is set to a zero matrix.

BLP-AC: Different from BLP-A, besides word alignments, Chinese WordNet synsets are used to build the intra-Chinese relations W_T, but the intra-English relations W_E is set to a zero matrix.

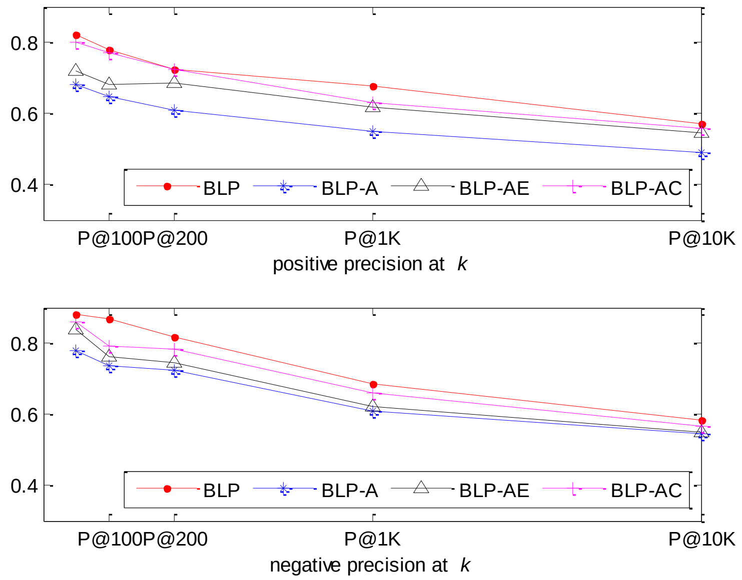


Figure 3.6 Evaluation on the intra-language relations (precision)

	Positive	Negative	Average
BLP-A	0.646	0.683	0.665
BLP-AE	0.670	0.685	0.678
BLP-AC	0.687	0.695	0.691
BLP	0.702	0.706	0.704

Table 3.4 Evaluation on the intra-language relations (recall)

As Figure 3.6 shows, when combining both the English and Chinese intra-language relations, the precision curves of both positive and negative predictions increase. It is clear that adding the intra-language relations has a positive influence on the BLP approach. This improvement in precision can be explained by the ability of the intra-language relations to refine the polarity scores. For example, the English word "sophisticated" is aligned to the positive Chinese word "精致的" (delicate) as well as the negative Chinese word "圆滑的" (wily, wicked). In the GI lexicon, the English word "sophisticated" is labeled as positive. When the bilingual graph includes only the inter-language relations, the negative Chinese word "圆滑的" is likely to be labeled as positive. However, with the intra-language relations, the negative Chinese word "圆滑的" may connect to the other negative Chinese words like "狡猾的" (foxy) and the Chinese positive word "精致的" may connect to the other positive Chinese words like "精巧的" (elaborate) because they are synonyms. Thus the polarity scores of the words are refined by the intra-language relations in each iteration of the propagation. Another advantage of the intra-language relations is that it reduces the noise brought by the inter-language relations. For example, sometimes the Chinese positive word "帮助" (help) is misaligned to the negative English word "freak" with the inter-language relation, but it is also connected to the synonyms "有助" (help) and "有益" (salutary) which are positive with the intra-language relations. The polarity score of the word "帮助" can be adjusted by the intra-language relations. Thus though the inter-language relations bring in certain noisy alignments, the intra-language relations help to refine the polarity scores of the words using their intra-language relations to a certain degree. As Table 3.4 shows, the recalls of these approaches also suggest that the proposed approach can achieve the further improvement in recall with the intra-language relations.

In summary, the combination of the English and Chinese intra-language relations improves the performance of the BLP approach. Combining with the conclusion from Section 3.2.2.2, it is confident to conclude that the inter-language relations help to discover more sentiment word candidates and the intra-language helps to refine the polarity scores of these candidates.

3.2.2.4 Influence of the Parameter $\boldsymbol{\beta}$

The parameter β in Equation E3.6 is to balance the proportions of inter-language propagation and intra-language propagation. This set of experiments aims to examine the influence of the parameter β by varying it from 0.1 to 1.

The precision curves with the change of β are shown in Figure 3.7. As observed, the precision curves (both positive and negative) tend to decline with the increase of β. As presented in Equation E3.6, a small β strengthens the intra-language propagation. Based on the conclusion from Section 3.2.2.3, intra-language propagation refines the polarity score and leads to an increase in precision. These two factors explain why we observe the high precisions at P@50 when β is around 0.1 and 0.2. With the increase of β, the effect of intra-language propagation is reduced gradually. Then the declination is observed in the curve of P@50. However, a large β enhances inter-language propagation through word alignments. The recall of the BLP approach then benefits from inter-language propagation because inter-language propagation helps to discover more sentiment words and consequently increase the recall. As a result, a slight increase is found in the curves of P@1K and P@10K with the increase of β. To balance inter-language and intra-language propagation, β is suggested to set at around 0.3.

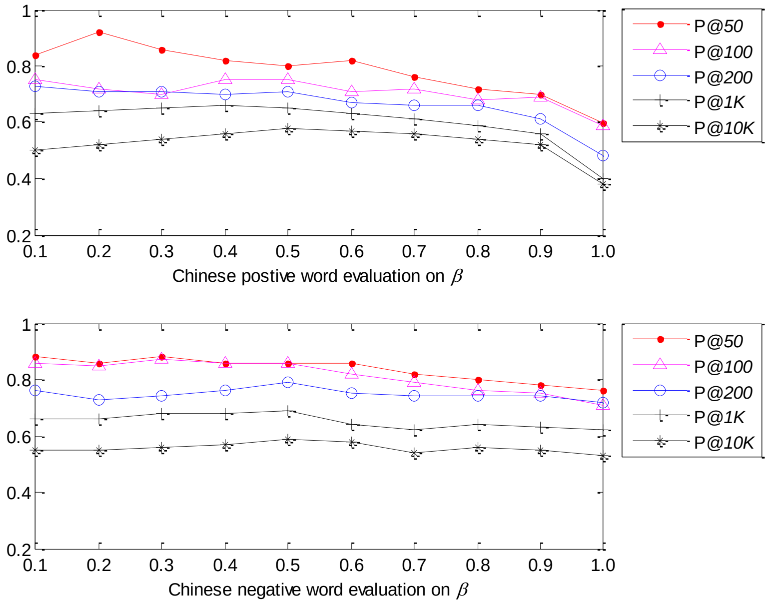


Figure 3.7 Influence of the parameter β

3.3 Signed Bilingual Graph based Sentiment Lexicon Learning

When the two words are connected by a relation in the bilingual graph introduced in Section 3.2, it assumes that these two words tend to share the same sentiment polarity. Based on this assumption, the proposed BLP approach transfers the sentiment information from one word to another along their connection. However, the word connections are more complex, not limited to the synonym or alignment relation. For example, if two words are connected by an antonym relation, they tend to own different sentiment polarities. The signed bilingual graph is constructed by incorporating the antonym relations into the bilingual graph. A signed bilingual graph label propagation approach is further proposed for cross-lingual sentiment lexicon learning based on the signed bilingual graph.

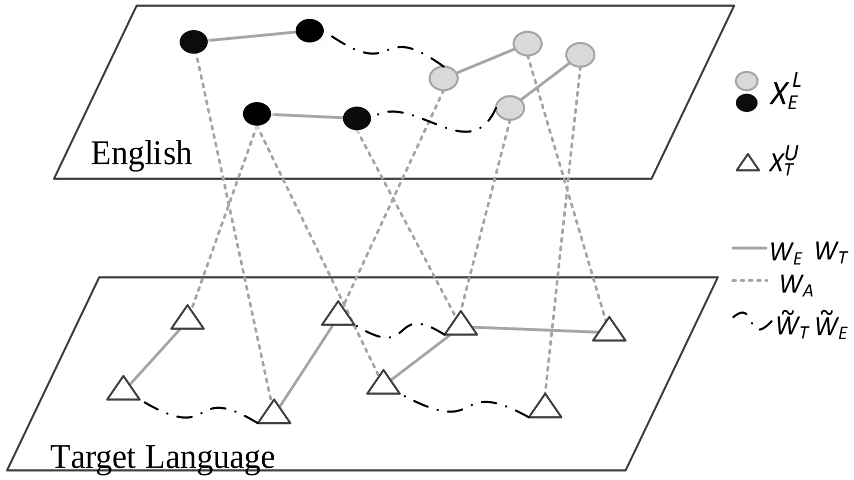


Figure 3.8 Signed bilingual graph for CSLL

3.3.1 Formalization of Signed Bilingual Graph based Sentiment Lexicon Learning

In Section 3.2.1, cross-lingual sentiment lexicon learning has been modeled by a bilingual graph $G = (X_E \cup X_T, W_E \cup W_T \cup W_A)$, in which the connected words are likely to share the same sentiment polarity. Beside synonyms, antonyms can also benefit sentiment lexicon learning. Let $\widetilde{W}_E$ and $\widetilde{W}_T$ denote the antonym relations in English and in the target language in Figure 3.8. The new bilingual word graph is then represented by $G = (X_E \cup X_T, W_E \cup W_T \cup W_A \cup \widetilde{W}_E \cup \widetilde{W}_T)$, where $W_E, W_T, W_A \in \mathbb{R}_+$ and $\widetilde{W}_E, \widetilde{W}_T \in \mathbb{R}_-$. Since this new bilingual graph contains both positive and negative connections, it is referred to as the *signed bilingual graph* [Leskovec et al. 2010]. The Signed Bilingual graph Label Propagation (**SBLP**) approach is proposed for cross-lingual sentiment lexicon learning based on the signed bilingual graph.

3.3.1.1 Signed Bilingual Graph Label Propagation

In the signed bilingual graph, the loss function $\Omega_l(F)$ and the smoothness function $\Omega_s(F)$ for synonym graph are similar to the Equation 3.2 and the Equation 3.3, while the antonym distance function $\widetilde{\Omega}(F)$ is defined as

$$\widetilde{\Omega}(F) = \frac{1}{2}\rho_3 \sum_{i,j=1}^{n} |\widetilde{w}_{E_{ij}}| \left\| \frac{f_{E_i}}{\sqrt{\tilde{d}_{E_{ii}}}} - \frac{f_{E_j}}{\sqrt{\tilde{d}_{E_{jj}}}} \right\| + \frac{1}{2}\rho_4 \sum_{i,j=1}^{m} |\widetilde{w}_{T_{ij}}| \left\| \frac{f_{T_i}}{\sqrt{\tilde{d}_{T_{ii}}}} - \frac{f_{T_j}}{\sqrt{\tilde{d}_{T_{jj}}}} \right\| \quad \text{E3.10}$$

If two words are connected by an antonym relation, the words should be as far as possible in the antonym graph. Unlike the Equation 3.1, the objective function for the antonym graphs is

$$argmax\left(\widetilde{\Omega}(F)\right) = argmax\left(\rho_3 \sum_{i,j=1}^{n} |\widetilde{w}_{E_{ij}}| \left\| \frac{f_{E_i}}{\sqrt{\tilde{d}_{E_{ii}}}} - \frac{f_{E_j}}{\sqrt{\tilde{d}_{E_{jj}}}} \right\| + \rho_4 \sum_{i,j=1}^{m} |\widetilde{w}_{T_{ij}}| \left\| \frac{f_{T_i}}{\sqrt{\tilde{d}_{T_{ii}}}} - \frac{f_{T_j}}{\sqrt{\tilde{d}_{T_{jj}}}} \right\|\right) \quad \text{E3.11}$$

Based on Equation 3.1 and Equation 3.11, the objective function of the signed bilingual graph label propagation is defined [Ma et al. 2009]

$$argmin(\Omega(F)) = argmin(\Omega_l(F) + \Omega_s(F) - \widetilde{\Omega}(F)) \quad \text{E3.12}$$

By differentiating the objective function of Equation 3.12 and setting it as zero, the optimal solutions are obtained in Equation 3.13.

$$\begin{aligned} F_E &= 2\mu(M_E - L_A M_T^{-1} L_A^T)^{-1}(Y_E - L_A M_T^{-1} Y_T) \\ F_T &= 2\mu(M_T - L_A^T M_E^{-1} L_A)^{-1}(Y_T - L_A^T M_E^{-1} Y_E) \end{aligned} \quad \text{E3.13}$$

where $M_E = 2(\rho_1 L_E - \rho_3 \tilde{L}_E + \mu I)$ and $M_T = 2(\rho_2 L_T - \rho_4 \tilde{L}_T + \mu I)$. $L_E = I - D_E^{-\frac{1}{2}} W_E D_E^{-\frac{1}{2}}$, $L_T = I - D_T^{-\frac{1}{2}} W_T D_T^{-\frac{1}{2}}$, $\tilde{L}_E = I - \widetilde{D}_E^{-\frac{1}{2}} \widetilde{W}_E \widetilde{D}_E^{-\frac{1}{2}}$, $\tilde{L}_T = I - \widetilde{D}_T^{-\frac{1}{2}} \widetilde{W}_T \widetilde{D}_T^{-\frac{1}{2}}$, and $L_A = I - D_{AL}^{-\frac{1}{2}} W_A D_{RL}^{-\frac{1}{2}}$. D_{AL} and D_{AR} are defined by Equation 3.6. Let $\tilde{L} = (I - \widetilde{D}^{-\frac{1}{2}} * \widetilde{W} * \widetilde{D}^{-\frac{1}{2}})$ and $\widetilde{D} = \sum_j |\widetilde{w}_{ij}|$. It has been proven by [Kunegis et al. 2010; Hsieh et al. 2012] that this signed graph Laplacian is positive semi-definite.

To obtain the optimal solutions in Equation 3.13, it requires to calculate the matrix inversion. Unfortunately, it is quite time-consuming to get the inversion especially when large numbers of vertexes are involved in a signed bilingual graph. Hence, the Jacobi algorithm [Saad 2003], which is a more efficient way to get the inversion is often used to obtain the solutions. The Jacobi algorithm is described as follows. Given the linear expression below

$$MX = b \tag{E3.14}$$

$M \in \mathbb{R}^{n \times n}$ is composed of a diagonal component D and a non-diagonal component Q. That is if

$$M = \begin{bmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{nn} \end{bmatrix} \tag{E3.15}$$

then

$$D = \begin{bmatrix} m_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & m_{nn} \end{bmatrix} \text{ and } Q = \begin{bmatrix} 0 & \cdots & m_{1n} \\ \vdots & 0 & \vdots \\ m_{n1} & \cdots & 0 \end{bmatrix} \tag{E3.16}$$

The Jacobi algorithm approximates the solution based on the iteration method below

$$X^{(t+1)} = D^{-1}(b - QX^{(t)}) \tag{E3.17}$$

The approximation can be also written as

$$x_i^{(t+1)} = \frac{1}{m_{ii}}\left(b_i - \sum_{j \neq i} m_{ij}\, x_i^{(t)}\right), \quad i = 1,2, \ldots, n \qquad \text{E3.18}$$

For cross-lingual sentiment lexicon learning, the iterative solutions in Equation 3.13 is obtained by replacing X with F_E and F_T and replacing M with $(M_E - L_A M_T^{-1} L_A^T)$ and $(M_T - L_A^T M_E^{-1} L_A)$, respectively. The iteration method is shown in Equation 3.19.

$$F_E^{(t+1)} = 2\mu \boldsymbol{Dg}(M_E - L_A M_T^{-1} L_A^T)^{-1}\left[Y_E - L_A M_T^{-1} Y_T - \boldsymbol{NDg}(M_E - L_A M_T^{-1} L_A^T) F_E^{(t)}\right]$$

$$F_T^{(t+1)} = 2\mu \boldsymbol{Dg}(M_T - L_A^T M_E^{-1} L_A)^{-1}[Y_T - L_A^T M_E^{-1} Y_E - \boldsymbol{NDg}(M_T - L_A^T M_E^{-1} L_A) F_T^{(t)}]$$

E3.19

where the functions $\boldsymbol{Dg}(M)$ and $\boldsymbol{NDg}(M)$ represent the diagonal and non-diagonal components of the matrix M. To perform the Jacobi algorithm, it needs to guarantee that the matrix M is positive semi-definite. The SBLP approach is summarized in Algorithm 3.2.

Algorithm 3.2 Signed Bilingual Graph Label Propagation

Input: Given $G = (X_E \cup X_T, W_E \cup \widetilde{W}_E \cup W_T \cup \widetilde{W}_T \cup W_A)$, the label Y_E for X_E. Initialize μ and $\rho_{1\sim4}$

Output: Label F_T for X_T

1. Initialize Y_E with the English sentiment seeds and set the Chinese initial sentiment label Y_T as zero
2. Calculate L_E, $\tilde{L}_E$, L_T, $\tilde{L}_T$ and L_A, then calculate M_E and M_T
3. Loop
4. Calculate F_E and F_T according to Equation 3.19
5. Until Y converge

3.3.2 Experiments

Again, the experiments are conducted on the ISI Chinese-English parallel corpus and the sentiment words from the GI lexicon are used as the sentiment seed words. The Chinese and English versions of WordNet are used to define the antonym word relations. These two lexicons provide the antonym pairs of synsets. From the two lexicons, totally 8,406 English and 6,312 Chinese antonym synset pairs are obtained. For a pair of synsets, any two words from the different synsets are regarded as the antonym to each other.

3.3.2.1 Evaluation on Signed Bilingual Graph

This set of experiments is to evaluate the role of the signed bilingual graph in cross-lingual sentiment lexicon learning. The SBLP approach is based on the signed bilingual graph $G = (X_E \cup X_T, W_E \cup W_T \cup W_A \cup \widetilde{W}_E \cup \widetilde{W}_T)$. The proposed SBLP approach is compared with the following two baselines.

MONO: This approach learns the Chinese sentiment lexicon based only on the Chinese monolingual graph $G_T = (X_T, W_T \cup \widetilde{W}_T)$. Since it needs the known sentiment information, the English labeled sentiment words X_E and the inter-language relations W_A are incorporated in the first iteration. Then X_E and W_A are set as zeros in later iterations.

SBLP-WOA: This approach is based on the synonym bilingual graph, which involves the inter-language relations W_A and the synonym intra-language relations W_E and W_T. $\widetilde{W}_E$ and $\widetilde{W}_T$ are set to zeros. In other words, the approach degenerates to the BLP approach.

Similar to [Zhou et al. 2003], μ is set to 0.1 in these approaches. Their precisions are shown in Figure 3.9. It shows that both the SBLP-WOA and SBLP approaches significantly outperform the Mono approach. The (signed) bilingual graph brings in more word relations and accelerate sentiment propagation. The English sentiment seed words are able to continuously provide the accurate sentiment information in the bilingual graph. Thus the increases of the SBLP-WOA and SBLP approaches in term of both precision and recall are observed in Table 3.5. Meanwhile, adding the antonym relations into the signed bilingual graph slightly improves the precision in the top ranked words. It appears that the antonym relations depict word relations in a more accurate way and refines the word sentiment scores more precisely. However, the synonym relations and the word alignment relations are dominating in the signed graph, while the antonym relations account for only a small percentage. It is hard for the antonym relations to introduce many new relations into the graph. Thus the antonym relations cannot help to further improve the recall.

Chinese	Positive	Negative	Average
MONO	0.641	0.649	0.645
SBLP-WOA	0.702	0.706	0.704
SBLP	0.708	0.709	0.709

Table 3.5 Evaluation on the MONO, SBLP-WOA and SBLP approaches (recall)

3.3.2.2 Sensitiveness of Parameters

ρ_1 and ρ_2 in Equation 3.3 tune the English and Chinese synonym intra-language propagation, respectively, while ρ_3 and ρ_4 in Equation 3.10 adjust the English and Chinese antonym intra-language propagation, respectively.

For simplicity, let ρ_1 equal to ρ_2 and ρ_3 equal to ρ_4. Then $\rho_{1,2}$ and $\rho_{3,4}$ are tuned together. When $\rho_{1,2}$ and $\rho_{3,4}$ range from {1e-1, 1, 10, 100, 1000}, Precision@1K ranges from 0.631 to 0.689 in average. In general, when $1 \le \rho_{3,4} < \rho_{1,2} \le 10$, the better results can be achieved.

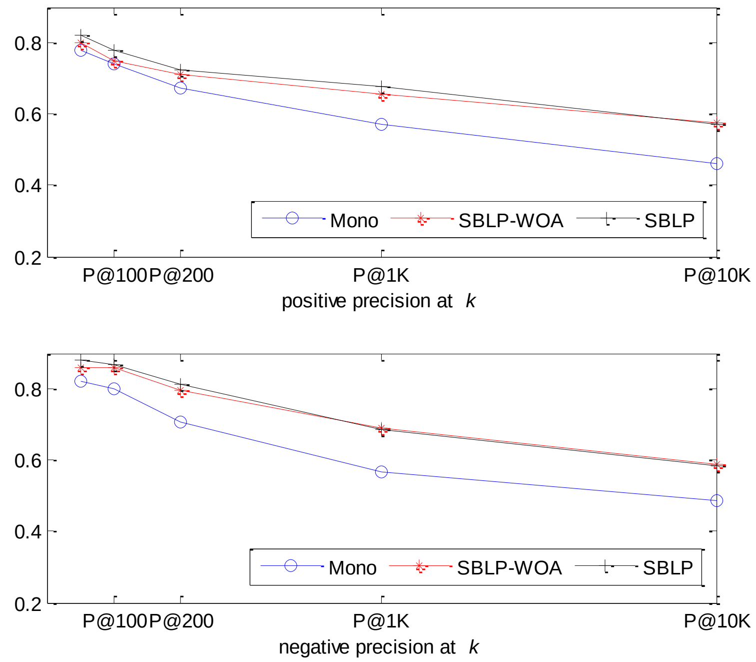


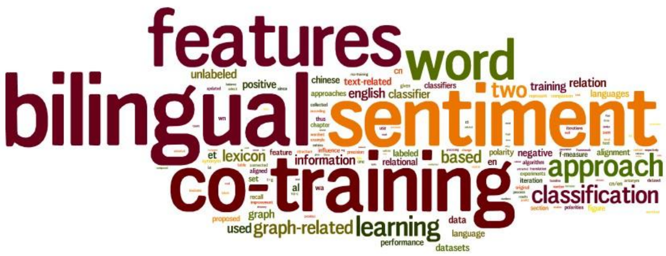
Figure 3.9 Evaluation on the MONO, SBLP-WOA and SBLP approaches (precision)

3.4 Chapter Summary

This chapter addresses cross-lingual sentiment lexicon learning based on bi-lingual graph, which incorporates the rich sentiment information from English to help sentiment lexicon learning for a non-English language (e.g., Chinese).

With respect to cross-lingual sentiment lexicon learning, the work here makes three breakthroughs. First, cross-lingual sentiment lexicon learning is formalized by the (signed) bilingual graphs. Two approaches are developed for cross-lingual sentiment lexicon learning and they both show advantages. Second, the word alignments derived from the parallel corpus are leveraged to connect the words in English and in the target language. It turns out that much more inter-language information is obtained from word alignments than from the bilingual dictionaries or from the machine translation results. The experimental results demonstrate that the performance based on word alignments remarkably improves the recall of the learned sentiment lexicon. Third, the antonym word relations are incorporated to form a signed bilingual graph. Based on the signed bilingual graph, a signed bilingual graph label propagation approach is presented. With the antonym relations, the SBLP approach achieves improvement in precision.

Chapter 4 Co-training based Bilingual Sentiment Lexicon Learning



4.1 Chapter Overview

With the bilingual graph based sentiment lexicon learning approaches introduced in Chapter 3, the sentiment polarity of a word is determined by the sentiment information of the connected words and the relations (e.g., synonyms or antonyms) between the connected words. This is one effective way to identify the sentiment polarity of a word. Apart from this, the information related to the word itself (e.g., the definition or the POS tag of the word) may also provide useful information to help infer its sentiment polarity.

In order to combine the two types of information, we approach sentiment lexicon learning as word level sentiment classification. A number of representative features concerning the two types of information have been explored. These features are categorized into the text-related features (e.g., the WordNet based features or the word alignment features) and the graph-related features (e.g., the numbers of connected positive words or negative words). The text-related features are extracted from the dictionaries or the text collections. They represent the syntactic and semantic information of the words. The text-related features are static and they never change during the learning process. Like in Chapter 3, the words in a language are connected by the synonym and antonym relations, naturally forming a word graph. The graph-related features are extracted from the word graph and to a certain degree they represent the structural information of the words in the word graph. Unlike the text-related features, the graph-related features dynamically change during the learning process when more unlabeled words become labeled.

With these features, we borrow the idea of co-training [Blum and Mitchell 1998] to address bilingual sentiment lexicon learning. We name this approach as *Co-training based Bilingual Sentiment Lexicon Learning* (**CBSLL**) since it is able to simultaneously expand or generate sentiment lexicons for two different languages under the co-training framework. The motivation behind CBSLL is that the co-training approach can make better use of the sentiment information from the two languages to mutually guide the sentiment lexicon learning processes of the two languages.

In the original co-training framework [Blum and Mitchell 1998], two groups of features are collected from the instances in the same dataset. The features in each group are assumed to be sufficient in indicating or identifying the classes of the instances. In each iteration of co-training, two classifiers predict the class of an unlabeled instance independently based on the different groups of features. If the instance is confidently predicted by either one of the two classifiers, it can be regarded as a new training instance and will be used to train the two classifiers in later iterations.

In our co-training based bilingual sentiment lexicon learning, a classifier is developed for each language to distinguish among the words with the positive, negative or neutral sentiment polarities. For a word, the corresponding features, i.e., the text-related and/or graph-related features are used to classify its sentiment polarity. If an unlabeled word in one language is confidently predicted as a sentiment polarity by the classifier, it not only will be regarded as a training word in the next iteration learning for the language, but also will provide one (or more) training word(s) for the other language through the bilingual word association. For example, the bilingual dictionaries can associate a word to its translation entry. We assume the two associated words are the same in sentiment polarity. If a word is a sentiment word with a sentiment polarity, the associated word(s) will have the same polarity. To verify the rationality of this assumption, we examine the sentiment polarity consistency of two associated words. It is observed that the top-associated words (for a word in one language, we rank its associated words in the other language according to their association frequencies) have the same sentiment polarity. It is thus reasonable to assign the sentiment polarity of a word to its top associated words. Through word association, the two learning processes in the co-training framework are bridged smartly.

A benefit of the co-training approach is that the sentiment information from the large amounts of unlabeled data can incrementally improve the classification performance. Specifically, at the end of each iteration, a certain number of most-confidence words in the unlabeled datasets are selected to update the training datasets in the next iteration. Meanwhile, the structure of the word graph changes when the unlabeled words get their sentiment polarities. For example, a word may become to connect more sentiment words than before. Thus, at the end of each iteration, we also need to update the graph-related features according to the changed word graph.

Again, English is selected as one of the languages in CBSLL because of its rich sentiment resources. It is easy to collect many sentiment words from the public available English sentiment lexicons, like SentiWordNet and MPQA, as the training data. Since the co-training process is advanced via the interactions between the sentiment information in the two languages, we are allowed to collect only a small set of training data for the other language and to transfer the rich sentiment information from English to the other language during the interactions. This will greatly reduce the effort on collecting the training data in the other language.

The rest of this chapter is organized as follows. Section 4.2 describes the text-related features and the graph-related features used in word level sentiment classification. In Section 4.3, a co-training based approach is presented to effectively integrate the static text-related features and the dynamic graph-related features. The experimental results of English and Chinese sentiment lexicon learning are discussed in Section 4.4. The chapter is concluded in Section 4.5.

4.2 Feature Selection for Bilingual Sentiment Lexicon Learning

For a classification task, one of the critical issues is to describe the instances with the representative features [Alpaydin 2010; Pang and Lee 2008]. This section is to describe the text-related and graph-related features used in word level sentiment classification.

4.2.1 Text-related Features

There are numbers of studies that work on feature selection for text classification [Forman 2003; Sebastiani 2002], sentence/document level sentiment classification [Pang et al. 2002] and information extraction [Riloff and Lehnert 1994; Téllez-Valero et al. 2005]. Relatively, it is more difficult to select the representative features for word level sentiment classification. Taking sentence/document level sentiment classification as examples, the words or phrases are the dominating features and are proven to be effective in classification [Pang et al. 2002; Dave et al. 2003]. Other specially-selected features such as sentiment words and emoticons can also be used to purse the further improvement. However, it is difficult to select features for a single word in word level sentiment classification because the word itself is often regarded as the minimal unit of sentiment polarity. In this section, the below features which are supposed to be helpful in word level sentiment classification are taken into consideration.

1. WordNet based Features

These features are extracted from WordNet. They have been widely used in the NLP tasks such as text classification [Amine and Mimoun 2007] and information retrieval [Rosso et al. 2004].

1) Part-Of-Speech (POS) features: The POS tag indicates the syntactic role of a word. It is the information commonly-used in many NLP researches [Mullen and Collier 2004]. Intuitively, some POS tags can tell whether a word is likely to be a sentiment word. For example, an adjective is likely to be a sentiment word. For CBSLL, five POS features are used. They are *Noun*, *Verb*, *Adverb*, *Adjective* and *Others.* A binary value (i.e., 0 or

1) is used to indicate if the word has the corresponding POS in WordNet. If the word is not included in WordNet, the "*Others*" feature is set to 1; otherwise it is 0.

2) Synonym feature: The synonymy relation connects the two words sharing the equivalent or similar meanings. The synonym relation is symmetrical. For example, the words "small & little", "big & large", "solon & statesman" are synonyms to each other. If a word has a positive (or negative) polarity, normally its synonyms are positive (or negative) as well. The synonyms of all the words in our dataset are used to construct the synonym feature space. For a word, a binary value (i.e., 0 or 1) is used to indicate if the word contains the corresponding synonym.

3) Antonym feature: The antonymy relation connects the two words holding the opposite or contradictory meanings. The antonymy relation is symmetrical as well. The example antonym pairs are "small & big", "fast & slow" and "young & old". Unlike the synonym, if a word is a positive (or negative) word, its antonyms are prone to be negative (or positive). The antonyms of all the words in our dataset make up the antonyms feature space. If a word has an antonym, the corresponding antonym feature is assigned to 1; otherwise, it is assigned to 0

4) Gloss feature: The gloss in WordNet is the definition of the synset, which can also be used as the approximate definition of the words in the synset [Szymański and Duch 2012]. For example, we can obtain three glosses for the word "beautiful" i.e., "delighting the senses or exciting intellectual or emotional admiration", "highly enjoyable" and

"aesthetically pleasing". Since the gloss is the definition of a word, the words in the same gloss tend to have the same sentiment polarity [Esuli and Sebastiani 2006]. We extract the glosses of all the words in our dataset and use the words in these glosses (after removing the stopwords and the low-frequency words) to form the gloss feature space. By doing so, a word can be represented by the words in its gloss(es) during classification.

2. Word Alignment Features

As mentioned in Chapter 3, the word alignment information is derived from the parallel corpus and conveys the word relations between a pair of parallel sentences (see Chapter 3 for more details). In some extent, the words aligned to a given word can be regarded as the synonyms of the given word in the other language. In our task, for the words in one language, all the aligned words in the other language are extracted. These aligned words are then used to construct the word alignment feature space. For a word, the weight of the word alignment features are initialized with the normalized alignment frequencies.

3. Translation Entry Features

Similar to the word alignment features, the translation entries in the bilingual dictionaries or from machine translators can be employed as the features of a word. Normally, the translation entries shall have the same sentiment polarity as the word. In our experiments, the translation entries are extracted from the LDC and UD dictionaries, from Google Translator or from Bing Translator[27]. The translation entries of all the words in our dataset are

[27] www.bing.com/translator

treated as the translation entry features. For a word, a binary value (i.e., 0 or 1) is used to indicate if the word can be mapped to the corresponding translation entry.

4.2.2 Graph-related Features

Recall that the words can form a word graph based on their relations like the synonym and antonym relations. The word graph contains certain amount of information that can indicate the sentiment polarities of the words. For example, for a given word, if its surrounding words (through synonym connections for example) are positive words, the given word tends to be a positive word as well. Motivated by this observation, some useful features are extracted from the word graph. We call them the *graph-related features*.

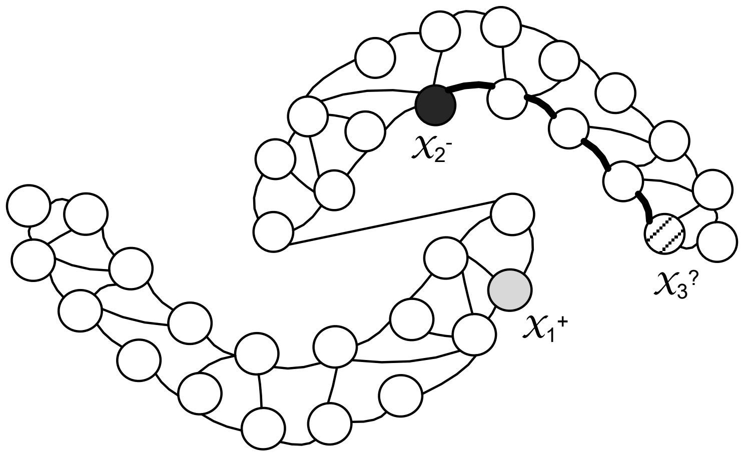


Figure 4.1 The two-moon distribution of the input data

There are two main advantages to explore the graph-related features. First, it has been reported in the past that the graph-related features can benefit to the learning problems [Camps-valls et al. 2007]. Figure 4.1 shows an example where the input distribution is the two-moon distribution [Zhu and Ghahramani 2002]. Let's assume the words x_1 and x_2 are the labeled positive and negative words. Now it is required to determine the sentiment polarity of the word x_3. Although the distance between the word x_1 and the word x_3 is closer than that between the word x_2 and the word x_3 according to the Euler distance, this, however, does not mean that the word x_1 and the word x_3 share the same sentiment polarity. Through the bold connections in the word graph, we can see that the word x_3 is actually more likely to share the same sentiment polarity with the word x_2. Without question, the graph information helps word sentiment classification. Second, during the learning process, when the unlabeled words obtain their sentiment labels, the word graph will gradually possess more sentiment information. From the evolving word graph with the increased sentiment information, we can explore more accurate features.

A signed graph [Leskovec et al. 2010] formed by the words and their relations have been introduced for sentiment lexicon learning in Chapter 3. In a certain degree, this signed graph is similar to a real online social network where a node can connect the other nodes positively or negatively, e.g., indicating *Trust* or *Distruct* in Epinions[28] and *Vote up* or *Vote down* in Digg[29]. Similarly, a word can connect to the other words according to the synonym or the antonym relations in the signed word graph. Learning from social network researches [Guha et al. 2004; Leskovec et al. 2010; Li et al. 2012], various graph characteristics are considered to define the graph-related features.

1. Structure Balance Features

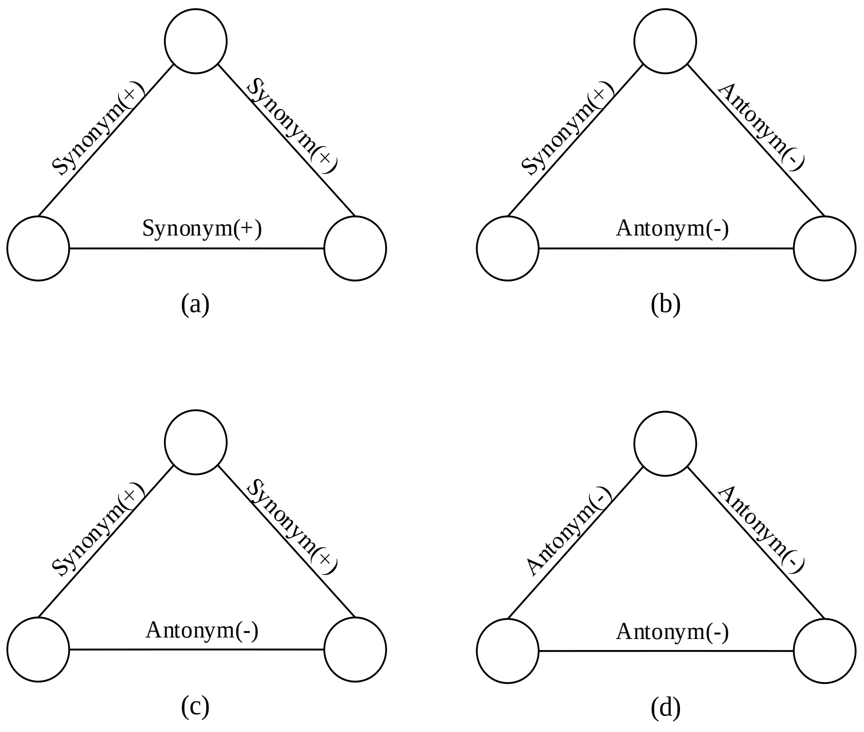
---

28 www.epinions.com

29 http://digg.com

The underlying theory of the structure balance stems from the research of social psychology in the late 1940s [Heider 1946]. Originally, the structure balance studies the relations between the people in a group. If two persons are friends, the connection between them is positive. Otherwise if they are foes, the connection is negative. The structure balance is then defined as:

Given three nodes in the graph (a node triangle), if the number of the three edges in the triangle is an odd number, the triangle is said to be balanced; otherwise, the triangle is said to be unbalanced. [Easley and Kleinberg 2010]



(a) and (b) are balanced triangles since the numbers of the positive connections are odd, while (c) and (d) are unbalanced triangles

Figure 4.2 Structure balance in sentiment lexicon learning

Figure 4.2 shows the application of the structure balance in sentiment lexicon learning. Like in Chapter 3, let the signs of the synonym connections be positive, and the signs of the antonym connections be negative. The nodes in the balanced triangles meet three requirements, i.e., for a word, "*the synonym of its synonym is its synonym*", "*the antonym of its synonym is its antonym*" and "*the antonym of its antonym is its synonym*".

The balanced/unbalanced graph structures in Figure 4.2 are represented by four graph-related features. For a given word, the numbers of the involved graphs are defined as the weights of the corresponding structure balance features.

2. Sentiment Polarity Endorsement Features

Intuitively, if a word connects to many positive (or negative) words through the synonym relations, its sentiment polarity tends to be positive (or negative) as well. On the contrary, if a word links to many positive (or negative) words through the antonym relations, its sentiment polarity is likely to be negative (or positive). According to this idea, six features are defined for a word to represent the numbers of the connected positive, negative and neutral words by the synonym relations and by the antonym relations, respectively.

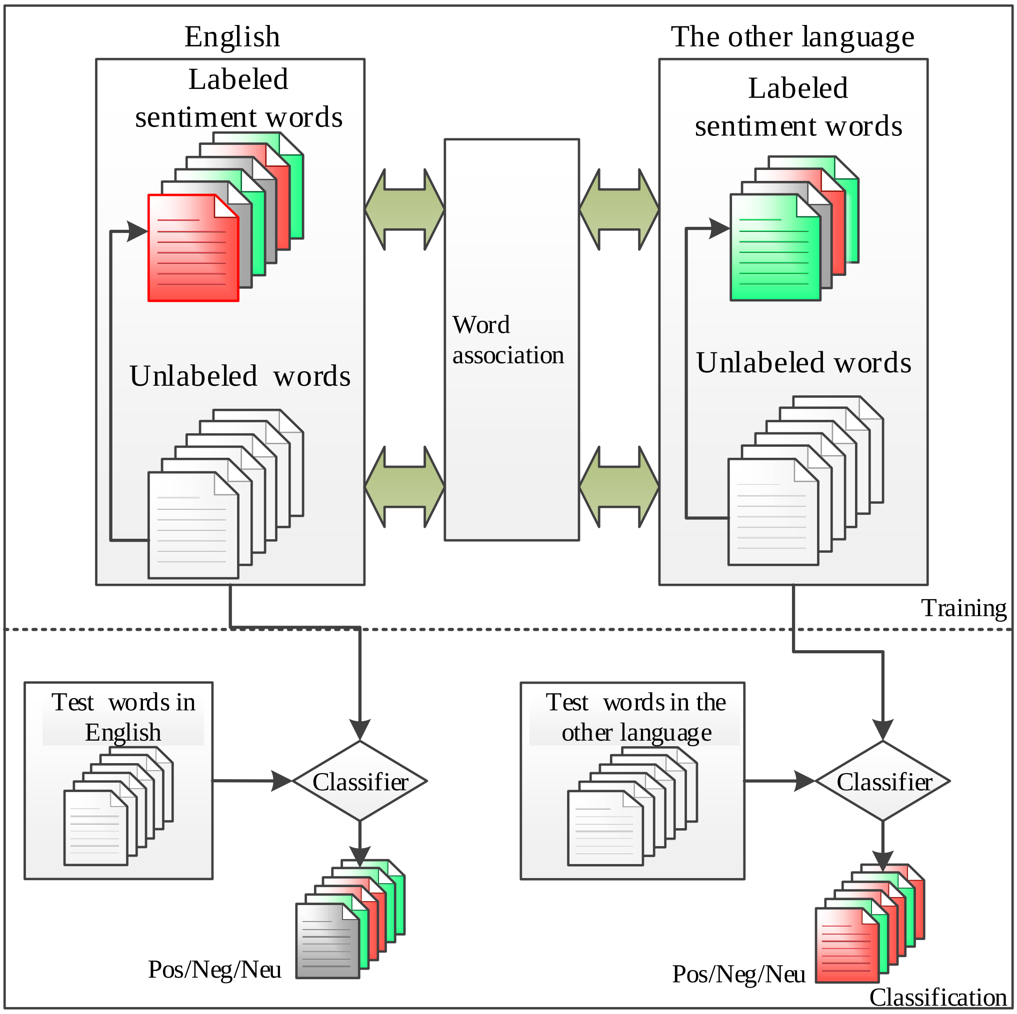


Figure 4.3 Co-training framework for bilingual sentiment lexicon learning

---

Algorithm 4.1 The co-training algorithm

Input: L_E, U_E and T_E represent the training, unlabeled and test data of English. L_T, U_T and T_T represent the training, unlabeled and test data of the other language. F_E and F_T are the features for the two languages, respectively. Initialize *p*, *q* and *N*

Output: Classifiers C_E and C_T

1. $k = 0$
2. **While** $k < N$ do
3. Train the classifier C_E on L_E with F_E
4. Train the classifier C_T on L_T with F_T

---

5. Use C_E to label the sentiment polarities of the words in U_E
6. Use C_T to label the sentiment polarities of the words in U_T
7. Select the top-p most-confident positive/negative/neutral words S_E from U_E
8. Select the top-p most-confident positive/negative/neutral words S_T from U_T
9. Select the top-q words A_T based on S_E through word association
10. Select the top-q words A_E based on S_T through word association
11. Remove S_E and A_E from U_E, and add them to L_E
12. Remove S_T and A_T from U_T, and add them to L_T
13. Update the graph-related features in F_E
14. Update the graph-related features in F_T
15. $k = k + 1$
16. **End While**
17. **Return** C_E and C_T

4.3 Co-training for Bilingual Sentiment Lexicon Learning

Using the text-related and graph-related features, a co-training based approach is proposed for bilingual sentiment lexicon learning in this chapter. The objective of the co-training approach is described as: given the sentiment (labeled) words and unlabeled words in two languages, the task is to develop two classifiers for the two languages, and then generate bilingual sentiment lexicons based on the results from the two classifiers. There are two main advantages of the proposed co-training approach. First, the sentiment information from the two languages can iteratively perform mutual-supervision in the learning processes of the two languages. Second, the proposed approach can effectively make use of the dynamic graph-related features. It allows the graph-related features to be updated in each iteration.

For each language, a set of sentiment words are collected to form the labeled dataset. Another set of words whose sentiment polarities are unknown are collected to form the unlabeled dataset. Using the word features extracted, two classifiers are developed based on the training and unlabeled datasets for the two languages, respectively. If an unlabeled word in one language is confidently predicted with a sentiment polarity, it is not only regarded as a new training word in the next iteration for the language, but also provides one (or more) training word for the other language through the bilingual word association. Here, we assume that the two associated words have the same sentiment polarity. For a confidently-predicted sentiment word in one language, we will assign the same sentiment polarity to its associated word(s) in the other language. Figure 4.3 illustrates the co-training framework for bilingual sentiment lexicon learning. It is divided into two phases, i.e., the training phase and the classification phase. In the training phase, the training words in English and in the other language are used to train the two classifiers, respectively. In each iteration, the classifiers predict the sentiment polarities of all words in the unlabeled datasets. The most-confident words and their associated words are selected and added to the corresponding training datasets for the next iteration. Meanwhile, the structure of the word graph changes when the unlabeled words obtain their sentiment polarities. Consequentially, the graph-related features change as well. Thus, after the unlabeled words obtain their sentiment polarities, the graph-related features are updated. Finally, the performance of the two classifiers is evaluated based on the predicted sentiment polarities of the words in the test datasets in the classification phase.

The proposed co-training approach is summarized in Algorithm 4.1. The algorithm is initialized with two sets of training words (i.e., L_E and L_T), two sets of unlabeled words (i.e., U_E and U_T) and two sets of test words (i.e., T_E and T_T). Two sets of features (i.e., F_E and F_T) are extracted for the training words in English and in the other language, respectively. In Line 3 and Line 4, two classifiers (i.e., C_E and C_T) are trained on the training words L_E and L_T with the features F_E and F_T, respectively. The classifiers are then used to predict the sentiment polarities of all the words in the unlabeled sets U_E and U_T in Line 5 and Line 6. Three lists of sentiment words (i.e., positive, negative and neutral) are generated from U_E and U_T for each language. The words in each list are ranked by their polarity scores. [Wan 2009] points out that a balanced growth of the training datasets for each class can ensure the performance of the co-training algorithm. Thus, in Line 7 and Line 8 the top-*p* ranked words in each list are selected as the most-confident words S_E and S_T. For each word in S_E and S_T, we rank its associated words according to the association information, for example, based on the associated frequencies in a corpus. Then the top-*q* associated ones are selected to form another two sets of words A_E and A_T in Line 9 and Line 10. These four sets of words (i.e., A_E, A_T, S_E and S_T) are used to expand the training datasets L_E and L_T in Line 11 and Line 12 respectively. In each iteration, the classifiers will predict the sentiment polarities of all the unlabeled words. According to these unlabeled words and their newly-predicted sentiment polarities, the graph-related features in F_E and in F_T are updated at the end of each iteration in Line 13 and Line 14. The algorithm stops when the maximum iteration limit is researched.

4.4 Experiments

The experiments are conducted on the following English and Chinese datasets. The sentiment words in the existing sentiment lexicons, *GI lexicon* and *OpinionFinder Lexicon*[30] [Wilson et al. 2005] are collected and used as the labeled English words. The sentiment words in these two lexicons are commonly used as the sentiment seed words in the previous work [Esuli and Sebastiani 2006; Mihalcea et al. 2007; Esuli and Sebastiani 2007; Baccianella et al. 2010]. If a word has two different sentiment labels in the two lexicons, the word is assumed to be neutral. From the two sentiment lexicons, 3,265 positive, 2,685 negative and 4,227 neutral English words are collected. In addition, 1,150 positive, 1,280 negative and 2,560 neutral Chinese words are manually annotated. For each language, the first two-third of the words in each class are used as the training data and the rest are used as the test data. The words in the ISI Chinese-English parallel corpus are used as the unlabeled dataset. *BerkeleyAligner* is run to align the parallel sentences in the ISI parallel corpus [Liang et al. 2006]. Based on the alignment information, the word alignment features are gathered.

Considering the performance of the two classifiers reflects the qualities of the generated bilingual sentiment lexicons, we first extensively evaluate the performance of the two classifiers on recognizing positive and negative words in the test datasets with the standard *precision*, *recall* and *F-measure*. We then evaluate the quality of the bilingual sentiment lexicons generated by the best-performance classifiers.

---

30 www.cs.pitt.edu/mpqa/subj_lexicon.html

4.4.1 Comparison with Baseline Approaches

This set of experiments is to evaluate the effectiveness of the proposed co-training approach, in which the word alignment is leveraged to bridge the English and Chinese learning processes. Specifically, if an unlabeled word is high-confidently predicted with a sentiment polarity by one classifier, all the words that are aligned to this word are sorted by the corresponding alignment frequencies and then the top-*q* aligned words are recommended to further improve the other classifier (in Line 9 and Line 10 in Algorithm 4.1). *p* is first set to be five analogous to [Wan 2009]. We will show the influence of *p* in Section 4.4.4. *q* is set to be one. We will show the influence of *q* in Section 4.4.5. The following two baseline approaches and the SBLP approach introduced in Chapter 3 are implemented for comparison.

SVM(EN/CN): In this approach, the supervised approach is employed for monolingual word sentiment classification. It uses the inductive SVM, *LibSVM*[31], for English/Chinese word sentiment classification and classification is based on the English/Chinese features. The English/Chinese training and test datasets are required but the English/Chinese unlabeled datasets are not used.

TSVM(EN/CN): In this approach, the semi-supervised approach is employed for monolingual word sentiment classification. It uses the transductive SVM, SVM^{light} [32], for English/Chinese word sentiment classification and classification is based on the English/Chinese features. Since SVM^{light} cannot perform the multi-class classification directly, like many other researchers, we

[31] www.csie.ntu.edu.tw/~cjlin/libsvm

[32] http://svmlight.joachims.org

employ the one-class-against-the-rest strategy [Joachims 2002]. The English/Chinese training, test and unlabeled datasets are needed in this approach.

SBLP: In the previous approaches as well as the co-training approach, we can use the learned classification models to predict words' sentiment polarities. These approaches can thus be called the model-based approaches. The SBLP approach proposed in Chapter 3 is different. It propagates the sentiment information among the word graph, and use the propagated sentiment scores to determine the sentiment polarities of the words. This process can be also considered as a process of word sentiment classification. Since the SBLP approach is mainly based on the word graph, we call it as the graph-based approach. In this set of experiments, the graph-based SBLP approach is provided for the comparison purpose. In particular, a bilingual graph is built with all the words in the training, test and unlabeled datasets in English and Chinese. The words in the English and Chinese training datasets are used as the sentiment seed words. The English and Chinese test datasets are used in evaluation.

The performances of these approaches are shown in Table 4.1. Three main findings are observed. First, the co-training and SBLP approaches significantly outperform the other approaches, which demonstrates the effectiveness of the co-training and SBLP approaches in word level sentiment classification. The main advantage of the co-training and SBLP approaches is that these two approaches can associate the sentiment lexicon learning of the two languages and leverage the sentiment information from the two languages to help sentiment lexicon learning of each language. That is why these two approaches perform well.

Second, in comparison with precision, the co-training approach gains more performance improvement in recall than the SBLP approach, especially for Chinese. In the SBLP approach, the sentiment polarity of a word is determined by the sentiment polarities of its connected words. If the word is not connected with the other words or is not extensively connected with the other words, the sentiment polarity of the word is hard to determine. In the co-training approach, apart from the sentiment information from the connected words, it can also use the word information to determine the sentiment polarity. Thus, the co-training approach is able to discover more sentiment words from different perspectives and leads to the better performance in recall. The significance of the co-training approach is that it allows us to include more information as features for sentiment classification.

Third, among the two baseline approaches, it is obvious that the transductive TSVM approach outperforms the inductive SVM approach. This finding suggests that the use of unlabeled data does benefit word level sentiment classification. Furthermore, the superior of the co-training approach to the TSVM approach demonstrates that the co-training approach is more suitable for making better use of the unlabeled datasets than the transductive SVM.

Based on these findings, we reach the conclusion that the co-training approach is effective in word level sentiment classification and has superiority in better use of the unlabeled datasets and in achieving higher recall.

Approach	Positive			Negative		
	Precision	Recall	F-measure	Precision	Recall	F-measure
SVM(CN)	0.700	0.549	0.615	0.681	0.543	0.604
TSVM(CN)	0.789	0.603	0.684	0.723	0.612	0.663
SBLP(CN)	0.817	0.612	0.700	0.781	0.655	0.712

Co-training(CN)	0.829	0.635	0.719	0.787	0.670	0.724
SVM(EN)	0.828	0.630	0.716	0.810	0.679	0.739
TSVM(EN)	0.857	0.636	0.730	0.811	0.702	0.753
SBLP(EN)	0.872	0.640	0.738	0.818	0.710	0.760
Co-training(EN)	0.869	0.659	0.749	0.823	0.719	0.768

Table 4.1 Comparison with the baseline approaches

4.4.2 Evaluation on Feature Selection

This set of experiments is to evaluate the effectiveness of the features used in word level sentiment classification.

Co-training(T): This approach uses the text-related features only.

Co-training(G): This approach uses the graph-related features only. We update the graph-related features at the end of each iteration.

The evaluation results of these approaches are presented in Table 4.2. It appears that the text-related features are more effective than the graph-related features. This is probability because that the number of the text-related features is significantly larger than that of the graph-related features. It is also easy to understand that the text-related features are more informative than the graph-related features. Thus it is not strange to see the text-related features outperform the graph-related features.

Approach	Positive			Negative		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Co-training(G)(CN)	0.728	0.604	0.660	0.739	0.600	0.663
Co-training(T)(CN)	0.812	0.633	0.711	0.780	0.640	0.703

Co-training(CN)	0.829	0.635	0.719	0.787	0.670	0.724
Co-training(G)(EN)	0.677	0.569	0.618	0.690	0.588	0.635
Co-training(T)(EN)	0.866	0.650	0.743	0.820	0.710	0.761
Co-training(EN)	0.869	0.659	0.749	0.823	0.719	0.768

Table 4.2 Evaluation on the text-related and graph-related features

As previously mentioned, the graph-related features are different from the text-related features. They are dynamically changed during the learning process. The following experiment is conducted to examine the effect of the graph-related feature update. Specifically, in the co-training(T+G-S) approach, though both the text-related and graph-related features are used, the graph-related features are not updated. As shown in Figure 4.4, without the update, the graph-related features do not influence the performance of the co-training(T+G-S) approach compared with the co-training(T) approach. When there is no update, the graph-related features, relative to the text-related features, are not sufficient for classification. It is found that the co-training approach with the graph-related feature update begins to outperform the co-training(T) and co-training(T+G-S) approaches after about the 50 iterations. With the increase of iterations, the graph contains more and more sentiment information. The graph-related features become denser and more accurate as well. When the graph-related features become informative enough, they will continuously play their roles in classification. From this set of experiments, we conclude that incorporating the continuously-evolving graph-related features in the co-training approach has the positive influence on word level sentiment classification.

From Figure 4.4, we have another two findings related to the influence of the iteration. First, the performance improvement of Chinese sentiment word classification in the first several iterations is greater and faster compared with that of English. The reason is that the initial size of the Chinese training dataset is smaller than that of the English training dataset. Thus the contribution of the newly complemented sentiment words is relatively greater and makes sentiment lexicon learning in Chinese faster. This also suggests that the proposed approach benefits a lot to sentiment lexicon learning of the non-English language, where the training data is rare or it is costly to obtain. The co-training approach has the ability to effectively transfer the rich sentiment information from English to the non-English language to help the lexicon learning process of the non-English language. Second, both the curves of the Chinese and English classification performance increase with the increase of the iteration number N. However, after they reach the peaks, the curves tend to decline. This means that the newly labeled words bring in more high-quality sentiment information at the beginning of the iterations. Thus, the performance increases. After a number of iterations, more newly labeled words are added into the training datasets and more noisy data are involved as well. The noisy data degrades the performance of prediction. To summarize, in general the co-training approach performs well with a wide range of iteration numbers. For a reasonable good performance, we would suggest to set the iteration number N between 90 and 150.

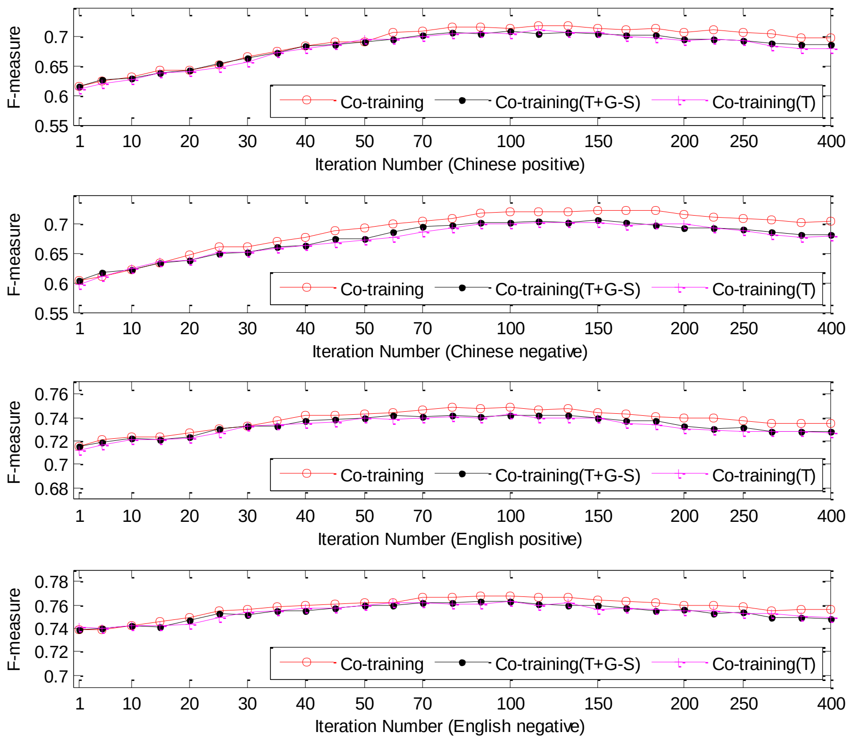


Figure 4.4 Influence of the update in the graph-related features.

4.4.3 Evaluation on Learning Processing Association

The co-training approach connects the words in the two languages to associate the learning processes of the two languages. In each iteration (Line 9 and Line 10 in Algorithm 4.1), for a high confidently predicted word in one language, its top-*q* associated words will be recommended for training the classifier in the other language. Three ways used to associate the words in the two languages have been introduced in Chapter 3. They are word alignment, bilingual dictionary and translation entry. The following experiments compare the three association methods.

Co-training (CN/EN)-A: This approach associates the learning processes with the word alignment derived from the parallel corpus. For a high confidently predicted sentiment word by one classifier, the aligned words are ranked according to their alignment frequencies and then the top-*q* ones are recommend for further training the other classifier.

Co-training (CN/EN)-D: This approach associates the learning processes with the bilingual dictionaries. For a high-confidently predicted sentiment word by one classifier, the translation entries from the two bilingual dictionaries, LDC and UD, are intersected. Since there is no way to rank these translation entries by intersection, the *q* entries are randomly selected for further training the other classifier.

Co-training (CN/EN)-M: This approach associates the learning processes with the machine translators. For a high confidently predicted sentiment word by one classifier, the average ranks of the returned translation entries from the Google Translator and from the Microsoft Translator are calculated. Then the top-*q* entries are recommended for further training the other classifier.

The experimental results are reported in Table 4.3. It shows only a slight difference among these approaches. This suggests that the top recommended words using the three association methods nearly have the same quality. In general, the word alignment approach is slightly superior to the other two in F-measure. Thus the word alignment is used to associate the two learning processes in later experiments.

4.4.4 Influence of the Parameter *p*

The parameter *p* controls the number of the newly labeled data being added to its own classifier for the next iteration of learning. The following experiments are conducted by varying *p* from 5 to 50. F-measure is shown in Figure 4.5. The F-measure curves of both positive and negative words in the test datasets decline when the value of *p* increases. As a matter of fact, the confidence of the newly-updated sentiment words declines when the value of *p* increases. The training datasets are likely to be updated with more noisy sentiment words. These noisy words influence the performance of the co-training approach and lead to the decreases. In this task, the preferred value of *p* is around 5.

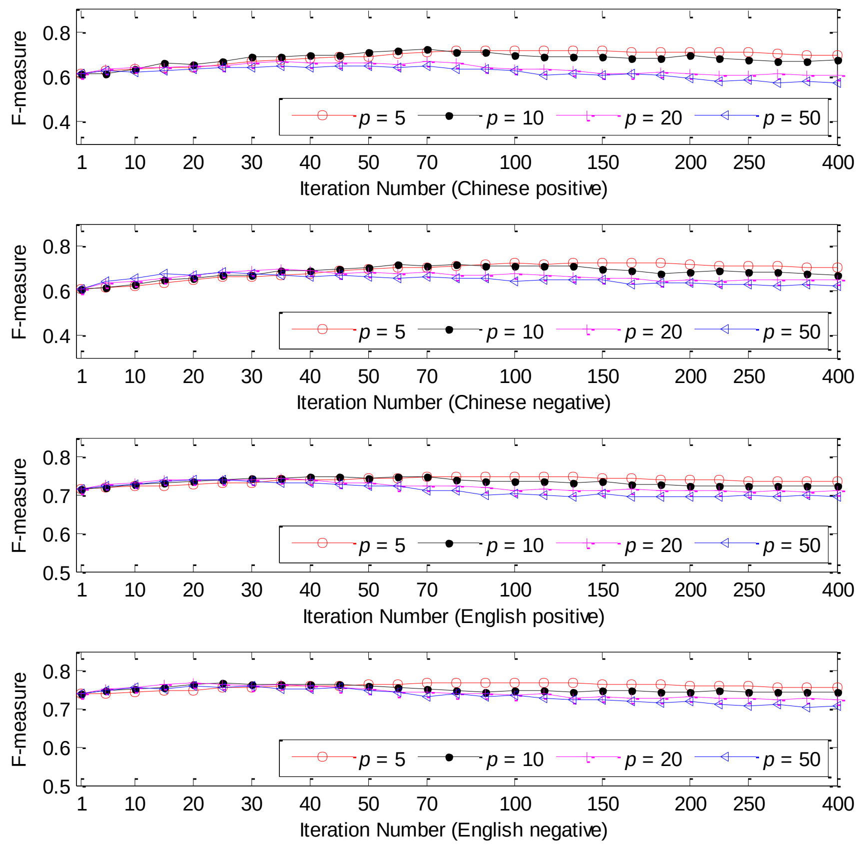


Figure 4.5 Influence of the parameter *p* in the co-training approach

Association approach	Positive			Negative		
	Precision	Recall	F-measure	Precision	Recall	F-measure

Co-training(CN)-A	0.829	0.635	0.719	0.787	0.670	0.724
Co-training(CN)-D	0.825	0.626	0.712	0.790	0.666	0.723
Co-training(CN)-T	0.823	0.621	0.708	0.780	0.662	0.716
Co-training(EN)-A	0.869	0.659	0.749	0.823	0.719	0.768
Co-training(EN)-D	0.886	0.647	0.748	0.829	0.704	0.761
Co-training(EN)-T	0.876	0.636	0.737	0.840	0.700	0.764

Table 4.3 Evaluation on the learning processing association methods

4.4.5 Influence of the Parameter *q*

The parameter q controls the number of the aligned words being added to the other classifier for the next iteration of learning. The following experiments are conducted by varying q from 1 to 10. In Figure 4.6, the performance is comparable when q is small (q = 1, 3). This verifies the rationality of our assumption that for a given word, its top-q aligned words can keep the consistent sentiment polarity. It is possible to assign the sentiment polarity of the given word to its top aligned words and use these aligned words to update the training dataset of the other classifier. Since the interaction of the two classifiers is based on the word alignment, this also ensures that in most cases the sentiment information can be correctly transferred between the languages. When q increases, the chances of the inconsistence of the sentiment polarities between the words and their aligned words increase. Certain aligned words with inconsistent sentiment polarities may be wrongly used to update the training datasets and consequently influence the performance of the co-training approach. Based on these experiments, it is more appropriate to set q less than 3.

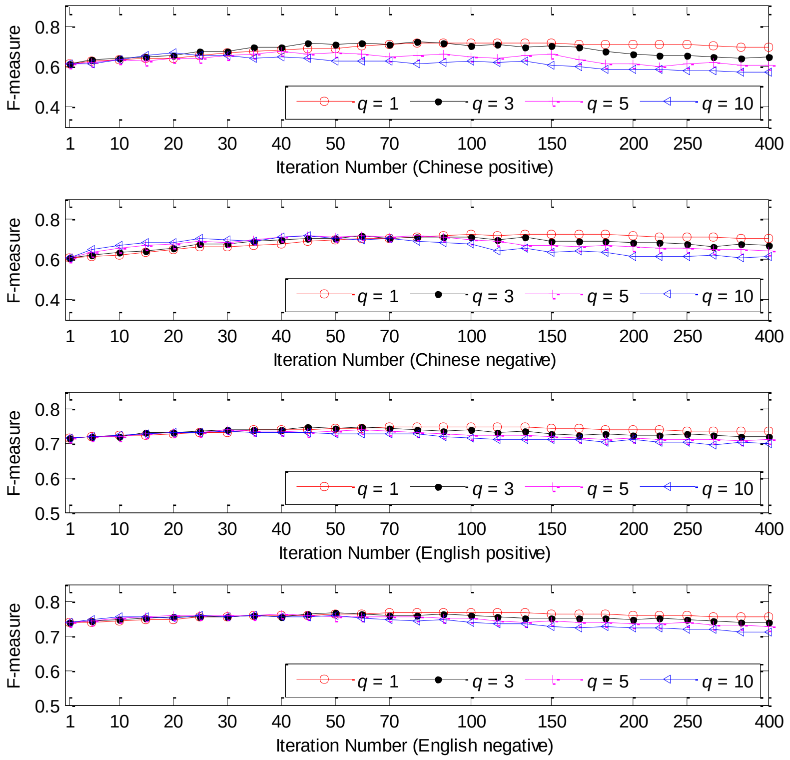


Figure 4.6 Influence of the parameter q in the co-training approach

4.4.6 Evaluation on Bilingual Sentiment Lexicons

The two trained classifiers (with the parameter settings of p, q and N as 5, 1 and 120) can then be used to generate bilingual sentiment lexicons by predicting the sentiment polarities of all the words in the unlabeled datasets. The words predicted to have the positive and negative classes are ranked according to the classifier outputs. Similar to the evaluation method used in Chapter 3, three annotators are invited to label the top-5000 ranked words in each approach. All the annotated words from each approach are included as the benchmark sentiment words for evaluation. The ranked lists are evaluated in term of

Precision@K and Recall. The following baselines and the graph-based SBLP approach are implemented for comparison.

Approach	Positive			Negative		
	P@100	P@1K	P@5K	P@100	P@1K	P@5K
SVM(CN)_L	0.710	0.542	0.483	0.690	0.516	0.465
TSVM(CN)_L	0.760	0.610	0.521	0.740	0.588	0.513
SBLP(CN)*	0.810	0.681	0.612	0.850	0.675	0.568
Co-training(CN)*	0.800	0.697	0.644	0.840	0.683	0.594
SVM(EN)_L	0.750	0.660	0.537	0.740	0.638	0.511
TSVM(EN)_L	0.770	0.681	0.564	0.760	0.651	0.521
SBLP(EN)*	0.820	0.744	0.591	0.840	0.712	0.573
Co-training(EN)*	0.830	0.762	0.607	0.830	0.710	0.593

Table 4.4 Evaluation on bilingual sentiment lexicons (precision)

SVM(EN/CN)_L: The classifiers from the baseline **SVM(EN/CN)** approaches are used to predict the sentiment polarities of the unlabeled (English/Chinese) words. The identified positive and negative sentiment words are ranked according to the classification outputs.

TSVM(EN/CN)_L: The classifiers from the baseline **TSVM(EN/CN)** approaches are used to predict the sentiment polarities of the unlabeled (English/Chinese) words. The identified positive and negative sentiment words are ranked according to the classification outputs.

SBLP: In the graph-based SBLP approach, the bilingual graph is built with the words in English and in Chinese. It then propagates the sentiment information from the sentiment seed words (words in the training datasets) to the unlabeled words. The identified positive and negative sentiment words are ranked according to the corresponding sentiment polarity scores outputted from the SBLP approach.

	Positive	Negative	Average
SVM(CN)_L	0.613	0.621	0.617
TSVM(CN)_L	0.651	0.663	0.657
SBLP(CN)	0.753	0.762	0.758
Co-training(CN)	0.808	0.811	0.809
SVM(EN)_L	0.685	0.703	0.694
TSVM(EN)_L	0.731	0.743	0.737
SBLP(EN)	0.782	0.788	0.785
Co-training(EN)	0.846	0.858	0.852

Table 4.5 Evaluation on bilingual sentiment lexicons (recall)

The evaluation results are shown in Table 4.4 and Table 4.5. We can find that the co-training and SBLP approaches significantly outperform the SVM and TSVM approaches, which indicates the effectiveness of the co-training and SBLP approaches in bilingual sentiment lexicon learning. As discussed in Section 4.4.1, compared with the SVM and TSVM approaches, the co-training and SBLP approaches can simultaneously leverage the sentiment information from the two languages in sentiment lexicon learning of each language. Thus the performances of these two approaches are better than the two baseline approaches. Furthermore, the co-training approach is superior to the SBLP approach in recall. This finding is consistent with what we have found in the experiments presented in Section 4.4.1, where the co-training approach also achieves a higher recall than the SBLP approach in the test dataset. As desired, the approach that has a good performance in word level sentiment classification should also perform well in bilingual sentiment lexicon learning. These results fit in with our expectation.

4.5 Chapter Summary

In this chapter, bilingual sentiment lexicon learning is cast as a problem of word level sentiment classification under the co-training framework. Both text-related features and graph-related features are explored from varieties of sources including WordNet, data collections and the word graph etc. A novel co-training based approach is presented to effectively integrate the text-related features and the graph-related features. Innovatively, when the unlabeled words obtain the sentiment polarities, the co-training approach allows the graph-related features to be updated. Based on the co-training approach, two classifiers are simultaneously developed for sentiment lexicon learning of the two languages, respectively. The evaluation on the classification performance demonstrates its advantages. The evaluation on bilingual sentiment lexicons further suggests that compared with the graph-based SBLP approach, the co-training approach learns more sentiment words and is superior in coverage.

Chapter 5 Multilingual Sentiment Lexicon Learning

5.1 Chapter Overview

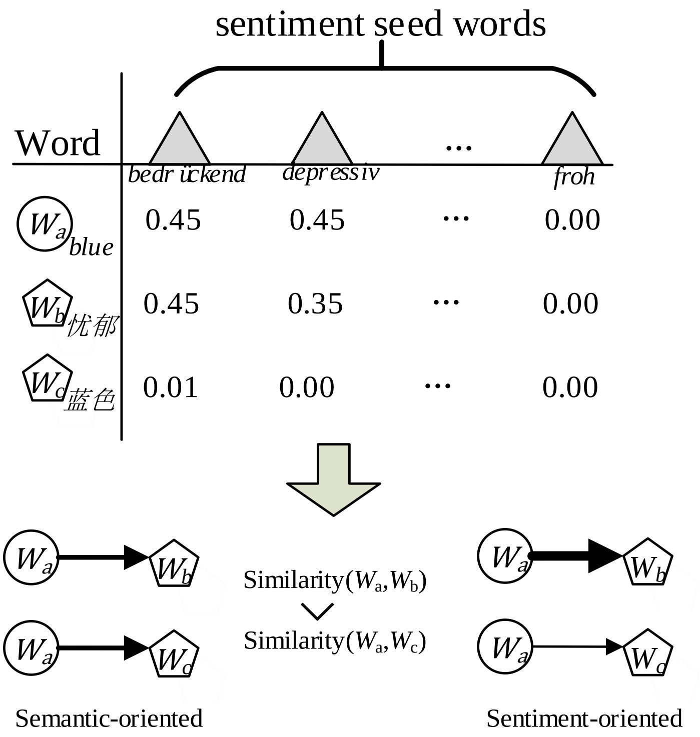
So far, I have discussed cross-lingual sentiment lexicon learning based on the bilingual graph or based on the co-training framework. In these two pieces of work, if two words are similar in semantics, they are connected or associated in sentiment lexicon learning. The underlying assumption is that the two words similar in semantics are also similar in sentiment. In fact, this assumption is not always valid especially when the words have multiple senses. Take the English word “blue” as an example. It corresponds to the translated Chinese words “蓝色” (the blue color) and “忧郁” (low in spirits) in Collins English-Chinese Dictionary[33]. When learning Chinese sentiment words, if “blue” is regarded as a negative sentiment word, the associations between them will transfer the negative sentiment information from “blue” to both “忧郁” and “蓝色”.

---

[33] www.collinsdictionary.com/dictionary/english

However, it is an irrefutable fact in Chinese that "忧郁" is a negative word, but "蓝色" is not. To alleviate this problem, the approach that can automatically recognize the sentiment-bearing connections for better sentiment propagation is expected. If a connection tends to link two sentiment words, I refer to this connection as a sentiment-bearing connection. Otherwise, it is called as a non-sentiment-bearing connection. The sentiment-oriented graph update is a process to update the original semantic-oriented connections towards the sentiment-bearing connections.

This chapter presents the work on *multilingual sentiment lexicon learning* (**MSLL**), where the multilingual (sentiment) resources are used not only to bring in more sentiment information, but also to guide the sentiment-oriented graph update. To this end, a novel *Sentiment-Oriented Label Propagation* (**SOLP**) approach is proposed. The SOLP approach propagates the sentiment information among multiple languages. It gradually refines the weights of the connections based on their sentiment-bearing degrees. The main idea of the sentiment-oriented graph update in SOLP is to use the benchmark sentiment words from the multiple languages to measure the sentiment-oriented consistency of the two words attached to a connection. The sentiment-oriented consistency of the two connected words describes how likely the two words are both sentiment words. Then I use the sentiment-oriented consistency to reweight the connections. Figure 5.1 illustrates the main idea of the sentiment-oriented graph update. Assume the word w_a is a sentiment word in one language. It is connected to the two words w_b and w_c in another language when w_a shares certain meanings, among which some are sentiment-bearing; some are not, with w_b and w_c. The SOLP approach attempts to identify the sentiment-bearing connections between w_a and w_b and/or between w_a and w_c and to use the appropriate one(s) to transfer the sentiment information. For this purpose, the benchmark sentiment words in the third language (or even more languages) are brought in. If w_a and w_b are both connected to the same third-language sentiment words, it implies that not only w_a and w_b are very likely to share the same or very similar semantic senses, but also the senses they share are very likely to be sentiment-bearing. In other words, both w_a and w_b tend to be sentiment words. By contrast, if w_c is rarely connected to the sentiment words, it seems that w_c may only share the non-sentiment-bearing senses with w_a. In this case, since w_b is more likely to be a sentiment word, the connection between w_a and w_b is more suitable for sentiment information propagation than that between w_a and w_c. Thus, the weight of the connection between w_a and w_b shall be increased, and meanwhile the weight of the connection between w_a and w_c shall be decreased. Refer back to the previous example where the English word “blue” is connected to the Chinese words “蓝色” and “忧郁”. Consider to use the German sentiment words as the guiding information. Since the words “blue” and “忧郁” tend to connect with the same German sentiment words, such as “bedrückend” (gloomy) and “depressiv” (depressed), the sentiment-oriented consistency between the words “blue” and “忧郁” is higher than that between the words “blue” and “蓝色”. Then the connections between “blue” and “忧郁” and between “blue” and “蓝色” are reweighted according to their degrees of sentiment-oriented consistency. In this way, the original semantic-oriented graph gradually shifts to the sentiment-oriented graph.



The words "bedrückend", "depressiv", "froh" are German words, which mean "gloomy", "depressed", and "happy", respectively.

Figure 5.1 Illustration of the sentiment-oriented graph update

The rest of this chapter is organized as follows. Multilingual sentiment lexicon learning is formalized with a semi-supervised problem based on a multilingual graph in Section 5.2. The sentiment-oriented label propagation approach is presented in Section 5.3. The experimental results on English, Chinese and German sentiment lexicon learning are discussed in Section 5.4. Next in Section 5.5, the qualities of the sentiment lexicons generated based on the different proposed approaches are further evaluated when they are applied to an application task, i.e., sentence level sentiment classification task. Section 5.6 concludes the chapter and suggests the future work on multilingual sentiment lexicon learning.

Algorithm 5.1: Sentiment-oriented Label Propagation

Input:

G multilingual graph

$\bigcup_{i=1}^{m} Z_i$ sentiment seed words from multiple languages

α_i $(i = \{1, \dots, m\})$ propagation parameters

μ_{ij} $(i, j = \{1, \dots, m\})$ balance parameters between sub-graphs

Output:

F final output sentiment label

1. Generate the initial sentiment label $Y = \{Y_1, \dots, Y_m\}$ and the initial similarity $A = \{A_1, \dots, A_m\}$ to sentiment seed words;
2. Generate the affinity matrixes S by normalizing the current weights W;
3. Propagate the sentiment label $F^{(t)}$;
4. Propagate the similarity $R^{(t)}$;
5. Based on $R^{(t)}$, perform the sentiment-oriented graph update for each sub-graph;
6. Loop line 2~5, until all the sentiment labels $F^{(t)}$ do not change;
7. Determine the sentiment polarities of the words by the propagated sentiment label F.

5.2 Formalization of Multilingual Sentiment Lexicon Learning

The objective of multilingual sentiment lexicon learning (MSLL) is to identify the sentiment polarities (either positive or negative) of the unlabeled words in multiple languages. A multilingual graph is set up by incorporating the words from multiple languages. With the sentiment seed words, multilingual sentiment lexicon learning is cast as a semi-supervised learning problem based on the multilingual graph.

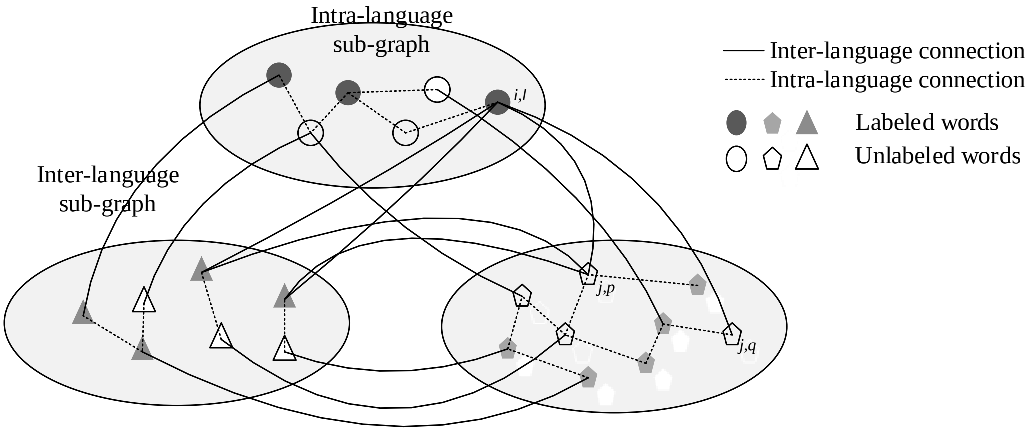


Figure 5.2 Multilingual graph for multilingual sentiment lexicon learning

Mathematically, let $X_i = \{x_{i1}, x_{i2}, \dots, x_{in_i}\}$ denote the words from the ith language and n_i represents the word number in the ith language. A word graph $G = (V, E, W)$ is built with the nodes $V = \bigcup_{i=1}^{m} X_i$ ($m \geq 3$). As illustrated in Figure 5.2, the multilingual graph is divided into two types of sub-graphs, including m intra-language sub-graphs, which contain the words from the same languages, and $m(m-1)$ inter-language sub-graphs, which are actually the bipartite graphs and connect the words in the two languages. $E_{i,j}$ denotes the edge from the word in the ith language to the word in the jth language. If $i = j$, $E_{i,j}$ is the intra-language connection. Otherwise it is the inter-language connection. W is the weight of the connection E and $W_{i,j}$ denotes the weight of the edge from the ith language to the jth language. $w_{ip,jq}$ denotes the weight of the edge from the word x_{ip} to the word x_{jq}. Each word in the multilingual graph is associated with a set of sentiment classes $C = \{c_1, \ldots, c_k\}$, where $k = 2$ referring to the positive and negative sentiment polarities in this work.

Let F denote the propagated sentiment label. The problem of multilingual sentiment lexicon learning can then be defined as: given the multilingual graph $\mathrm{G} = (V, E, W)$ and the initial sets of sentiment seed words $Z = \bigcup_{i=1}^{m} Z_i$ (i.e., the labeled words) from m languages, MSLL is to compute f_{ij} for all the unlabeled words x_{ij} $(x_{ij} \notin Z)$. Based on f_{ij}, the sentiment polarity of the unlabeled word x_{ij} is determined.

5.3 Sentiment-oriented Label Propagation

The sentiment-orientated label propagation (SOLP) approach is proposed to simultaneously transfer the sentiment information among multiple languages and adjust the weights of the word connections. Algorithm 5.1 describes the main framework of the sentiment-oriented label propagation approach where the SOLP approach is divided into three main components, *multilingual sentiment information propagation* (Lines 2 ~ 4), *sentiment-oriented graph update* (Line 5) and *sentiment polarity identification* (Line 7). The illustration of the SOLP approach is shown in Figure 5.3. Multilingual sentiment information propagation is to propagate the sentiment information among the multilingual graph. Sentiment-oriented graph update is to update the original semantic-oriented connections to the sentiment-oriented connections. In this stage, the similarities of a word to all the multilingual sentiment seed words are computed. Similar to sentiment information propagation, this similarity information is also allowed to propagate among the multilingual graph. In each iteration, we use the similarities to calculate the sentiment-oriented consistencies of the two connected words and use these sentiment-oriented consistencies to update the graph weights. Sentiment polarity identification is to determine the sentiment polarities of the words by the propagated sentiment labels. Each component will be presented in detail below.

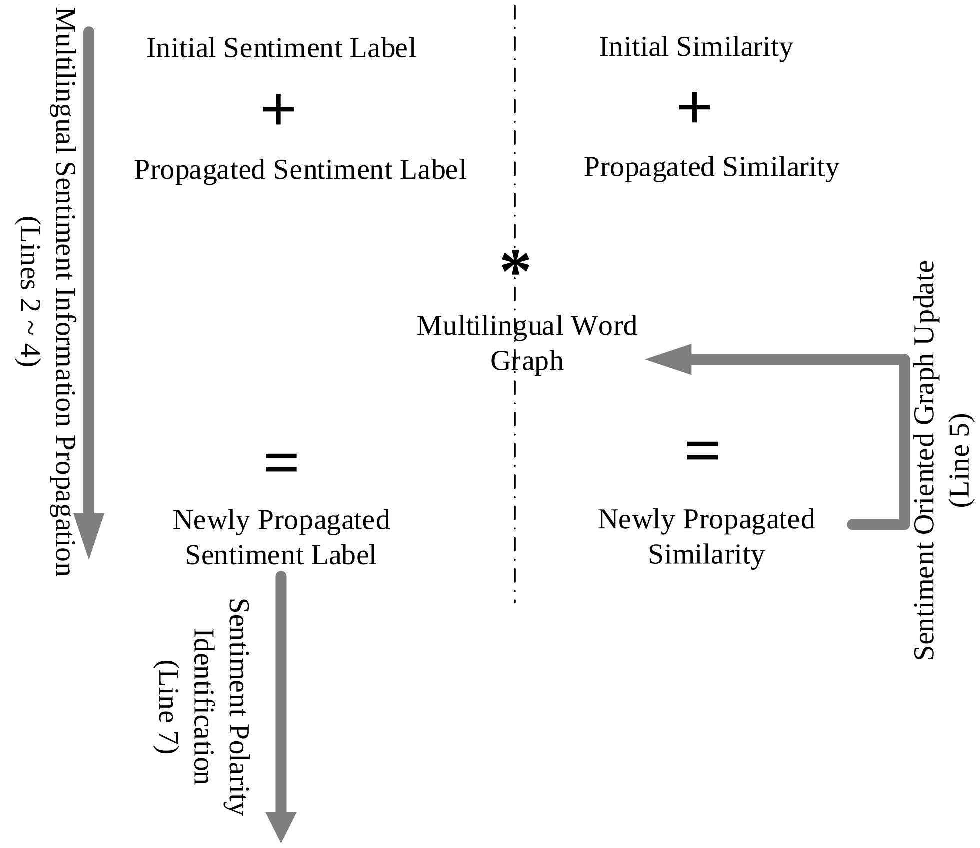


Figure 5.3 Iteration illustration of the SOLP approach

5.3.1 Multilingual Sentiment Information Propagation

The multilingual graph contains two types of sub-graphs, the intra-language sub-graphs and the inter-language sub-graphs. The smoothness function of the intra-language sub-graphs is

$$\Omega^{intra}(F_{**k}) = \sum_{i=1}^{m} \mu_{ii} \sum_{p,q=1}^{n_i} w_{ip,iq} \left(\frac{f_{ipk}}{\sqrt{d_{ip,ip}}} - \frac{f_{iqk}}{\sqrt{d_{iq,iq}}} \right)^2 \qquad \text{E5.1}$$

where the star $(*)$ denotes all the elements. For example, F_{i*k} denotes the propagated sentiment labels of all the words in the ith language on the class c_k. f_{ipk} denotes the propagated sentiment label of the word x_{ip} on the class c_k. For the ith language, $D_{i,i}$ is defined as an $n_i \times n_i$ diagonal matrix. The (p,p) element of $D_{i,i}$ (i.e., $d_{ip,ip}$) is the sum of the pth row of $W_{i,i}$. μ $(0 \leq \mu_{ii} < 1)$ are the balance parameters among different languages. If the affinity matrix of $W_{i,i}$ is defined as $S_{i,i} = (D_{i,i})^{(-1/2)} W_{i,i} (D_{i,i})^{(-1/2)}$, the smoothness function of the intra-language sub-graphs in Equation 5.1 can be rewritten as

$$\Omega^{intra}(F_{**k}) = 2 \sum_{i=1}^{m} \mu_{ii} F_{i*k}^{T} \left(I_{i,i} - S_{i,i} \right) F_{i*k} \qquad \text{E5.2}$$

where $I_{i,i}$ is the identity matrix of size $n_i \times n_i$.

The inter-language sub-graphs are bipartite graphs and the smoothness function of the inter-language sub-graphs is defined as

$$\Omega^{inter}(F_{**k}) = \sum_{\substack{i,j=1 \\ (i \neq j)}}^{m} \mu_{ij} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} w_{ip,jq} \left(\frac{f_{ipk}}{\sqrt{d_{ip,ip}^{L}}} - \frac{f_{jqk}}{\sqrt{d_{jq,jq}^{R}}} \right)^2 \qquad \text{E5.3}$$

$D_{i,i}^{L}$ is defined as an $n_i \times n_i$ diagonal matrix, of which the (p,p) element $d_{ip,ip}^{L}$ is the sum of the p row of $W_{i,j}$. $D_{j,j}^{R}$ is defined as an $n_j \times n_j$ diagonal matrix, of which the (q,q) element $d_{jq,jq}^{R}$ is the sum of the q columns of $W_{i,j}$. The affinity matrix of $W_{i,j}$ is then defined as $S_{i,j} = (D_{i,i}^{L})^{(-1/2)} W_{i,j} (D_{j,j}^{R})^{(-1/2)}$. A matrix $L_{i+j,i+j}$ is employed to bind $S_{i,j}$ and $S_{j,i}$

$$L_{i+j,i+j} = \begin{bmatrix} 0 & S_{i,j} \\ S_{j,i} & 0 \end{bmatrix}$$

The smoothness function of the inter-language sub-graphs in Equation 5.3 is then rewritten as

$$\Omega^{inter}(F_{**k}) = \sum_{i,j=1(i\neq j)}^{m} \mu_{ij} F_{(i+j)*k}^{T} (I_{i+j,i+j} - L_{i+j,i+j}) F_{(i+j)*k} \qquad \text{E5.4}$$

and $F_{(i+j)*k} = [F_{i*k}^T, F_{j*k}^T]^T$. Each μ_{ij} is a value between 0 and 1 and $\sum_{j=1}^{m} \mu_{ij} = 1$.

Similar to the label propagation approach in [Zhou et al. 2003], we also define the loss function

$$\begin{aligned}\Omega^{fitting}(F_{**k}) &= \sum_{i=1}^{m} \alpha_i \sum_{p=1}^{n_i} \left(f_{ipk} - y_{ipk}\right)^2 \\ &= \sum_{i=1}^{m} \alpha_i (F_{i*k} - Y_{i*k})^T (F_{i*k} - Y_{i*k})\end{aligned} \qquad \text{E5.5}$$

The objective function for MSLL based on the multilingual graph is the summation of all the above smoothness functions and the loss function, i.e.,

$$\Omega(F_{**k}) = \Omega^{intra}(F_{**k}) + \Omega^{inter}(F_{**k}) + \Omega^{fitting}(F_{**k}) \qquad \text{E5.6}$$

By differentiating Equation 5.6 with respect to each F_{i*k} and letting $\frac{\mathrm{d}(\Omega(F_{**k}))}{\mathrm{d}(F_{i*k})}$=0 for all languages, the closed form solution can be obtained, i.e.,

$$F_{i*k} = P^{-1} \left(\alpha_i Y_{i*k} + \sum_{j=1, j\neq i}^{m} \mu_{ij} S_{i,j} F_{j*k} \right) \qquad \text{E5.7}$$

and $P = (2\mu_{ii} + \alpha_i) I_{i,i} + \sum_{j=1,\mathrm{j}\neq \mathrm{i}}^{m} \mu_{ij} S_{i,j}$

Y_{i*k} denotes the initial sentiment labels of the words in the ith language on the class c_k. According to [Ji et al. 2010], Equation 5.8 is proven to converge to the close form solution in Equation 5.7.

$$F_{i*k}^{(t)} = \frac{\sum_{j=1}^{m} \mu_{ij} S_{ij} F_{j*k}^{(t-1)} + \mu_{ii} S_{ii} F_{i*k}^{(t-1)} + \alpha_i Y_{i*k}}{\mu_{ii} + \alpha_i + 1} \qquad \text{E5.8}$$

5.3.2 Sentiment-oriented Graph Update

As shown in Figure 5.3, the purpose of the sentiment-oriented graph update is to gradually update the original word connections to the sentiment-bearing connections. In the SOLP approach, the word similarities to the sentiment seed words $\bigcup_{i=1}^{m} Z_i$ are also propagated in the multilingual graph. These similarities are used to calculate the sentiment-oriented consistencies of the two connected words. The sentiment-oriented consistencies are then used to update the connections.

Let z_{ij} denote the jth sentiment word from the ith language. R denotes the propagated similarities to the sentiment seed words, and r_{iplk} denotes the similarity of the word x_{ip} to the sentiment seed word z_{lk}. Similar to Equation 5.8, the propagated similarities R_{i*lk} can be iteratively calculated according to

$$R_{i*lk}^{(t)} = \frac{\sum_{\text{j}=1}^{m} \mu_{ij} S_{ij} R_{j*lk}^{(t-1)} + \mu_{ii} S_{ii}\, R_{i*lk}^{(t-1)} + \alpha_i A_{i*lk}}{\mu_{ii} + \alpha_i + 1} \qquad \text{E5.9}$$

The <u>*S*</u>*entiment-oriented* <u>*Con*</u>*sistency* (***SCon***) of the two words describes how likely they are both sentiment words. It is calculated by the cosine measure between the similarity vectors of the two connected words. For example, $SCon$ between the word x_{ip} and the word x_{jq} for the l th language is computed by

$$SCon_l(x_{ip}, x_{jq}) = \frac{< r_{ipl*} \cdot r_{ipl*} >}{\sqrt{< r_{ipl*} \cdot r_{ipl*} > * < r_{ipl*} \cdot r_{ipl*} >}} \qquad \text{E5.10}$$

where $< r_{ipl*} \cdot r_{ipl*} >$ denotes the inner product of the similarities r_{ipl*} and r_{ipl*}. In each iteration, the connection $w_{ip,jq}$ is refined with the normalized sentiment oriented consistency in all the languages except the i, j languages, i.e.,

$$w_{ip,jq}^{(t)} = w_{ip,jq}^{(t-1)} \sum_{l=1,(l\neq i,j)}^{m} \frac{\gamma_l (SCon_l^{(t-1)}(x_{ip}, x_{jq}) + \sigma)}{\sqrt{\sum_{r=1}^{n_j} (SCon_l^{(t-1)}(x_{ip}, x_{jr}))^2}} \quad \text{E5.11}$$

where $\sum_{l=1}^{m} \gamma_l = 1$ and they are used to balance the consistencies among the sentiment seed words from the different languages. σ is a positive smooth parameter to avoid the weights dropping to 0.

5.3.3 Sentiment Polarity Identification

Sentiment identification is the final stage to determine the sentiment polarities of the unlabeled words. If $x_{ip} \in Z_i$ is a positive sentiment seed word, the sentiment label y_{ip1} is initialized as one and y_{ip2} is set to zero. If it is a negative sentiment seed word, y_{ip1} is initialized as zero and y_{ip2} is set to one. For the unlabeled word $x_{jq} \notin Z_j$, y_{jq1} and y_{jq2} are both set to zeros. The proposed SOLP approach will generate the sentiment label F for each unlabeled word. For the unlabeled word $x_{jq} \notin Z_j$, if $(f_{jq1} - f_{jq2}) > \varepsilon$, it is determined as a positive word. If $(f_{jq2} - f_{jq1}) > \varepsilon$, it is judged as a negative one. Otherwise, the word is assumed to be a neutral word. Each word has m sets of similarities to the sentiment seed words from m languages. If $w_{ip,jq}$ is non-zero, the initial similarity a_{ipjq} is initialized with the row normalized $w_{ip,jq}$. The propagated similarities R are used to calculate the sentiment-oriented consistencies and to guide the graph update, but not to identify the sentiment polarities of the unlabeled words.

5.4 Experiments

In the following experiments, the words in English (EN), Chinese (CN) and

German (GE) are used to build the multilingual graph. The synonym relations are employed to set up the intra-language relations. The words in the different languages are connected by the word alignments derived from the parallel corpus called MultiUN (version2), which contains the official documents of the United Nations [Chen and Eisele 2012]. The words that appear more than ten times in the corpus are included in the multilingual graph. The parallel sentences are aligned using the *BerkeleyAligner* aligner. The weights of the inter-language relations are initialized by the word alignment frequencies. For the intra-language sub-graph, two words are connected if they are synonyms and the weight on the connection is set to one. The synonym relations from English, Chinese and German (GermaNet[34]) WordNets are used to build the intra-English/Chinese/German sub-graphs. The details of the experimental data are provided in Table 5.1. For the sentiment seed words, 2,005 positive and 1,635 negative English words from the GI lexicon are used as the English sentiment seed words. 1,650 positive and 1,808 negative German words are used as the German sentiment seed words [Remus et al. 2010; Wiebe and Riloff 2005].

Language	Words	Inter-language relation	Intra-language relation
English	32.1K	76.0K	20.7K
Chinese	54.4K	119.5K	17.7K
German	17.5K	48.4K	5.6K

Table 5.1 Experiment data from MultiUN (v2)

Based on SOLP, the predicted positive and negative words are sorted

[34] www.sfs.uni-tuebingen.de/lsd

according to their corresponding sentiment scores. Similar to the evaluation in TREC Blog Distillation [Ounis et al. 2010], three annotators are invited to label the top 5000 ranked words in each approach. All the annotated words from each approach are included as the benchmark sentiment words for evaluation. The ranked lists are evaluated in term of Precision@K and Recall. Since we are familiar with Chinese, in the following we will first focus on the evaluation of the learned Chinese sentiment lexicon. The evaluation on the learned English and German sentiment lexicons will be provided in Section 5.4.5.

5.4.1 Evaluation on Multilingual Resources

This set of experiments is to examine the role of multilingual resources in sentiment lexicon learning. In order to examine the usefulness of multilingual resources, for the time being, the update of the multilingual graph is ignored. MLP (Multilingual Label Propagation) is used to represent the SOLP approach when there is no sentiment-oriented graph update. Without the sentiment-oriented graph update, the MLP approach can be regarded as an extension of the SBLP approach introduced in Chapter 3, which extends from the bilingual graph to the multilingual graph. The MLP approach compares with the following baseline approach.

Baseline: This approach assumes that a word tends to be positive (or negative) if it connects to more positive (or negative) words. For a given word x_{ip} in the multilingual graph, the summations of the weights, respectively to the positive and negative seed words, are used to identify the sentiment polarity of the word.

$$Pos(x_{ip}) = \sum_{j=1}^{m} \sum_{q=1}^{|Z_j^+|} w_{ip,jq} \text{ and } Neg(x_{ip}) = \sum_{j=1}^{m} \sum_{q=1}^{|Z_j^-|} w_{ip,jq} \quad \text{E5.12}$$

where Z_j^+ denotes the positive sentiment seed words in the jth language and Z_j^- denotes the negative ones. If $(Pos(x_{ip}) - Neg(x_{ip})) > \varepsilon$, the word x_{ip} is determined as a positive word. If $(Neg(x_{ip}) - Pos(x_{ip})) > \varepsilon$, the word x_{ip} is determined as a negative word. Otherwise, the word is considered to be a neutral word.

MLP(Tri) and Baseline(Tri) mean to incorporate multilingual (Chinese, English and German) resources in sentiment lexicon learning. For comparison, MLP(Cn-En) and Baseline(Cn-En) represent the incorporation of bilingual (Chinese and English) resources in the learning process. Similarly, MLP(Cn-Ge) and Baseline(Cn-Ge) represent the incorporation of the Chinese and German resources in the learning process.

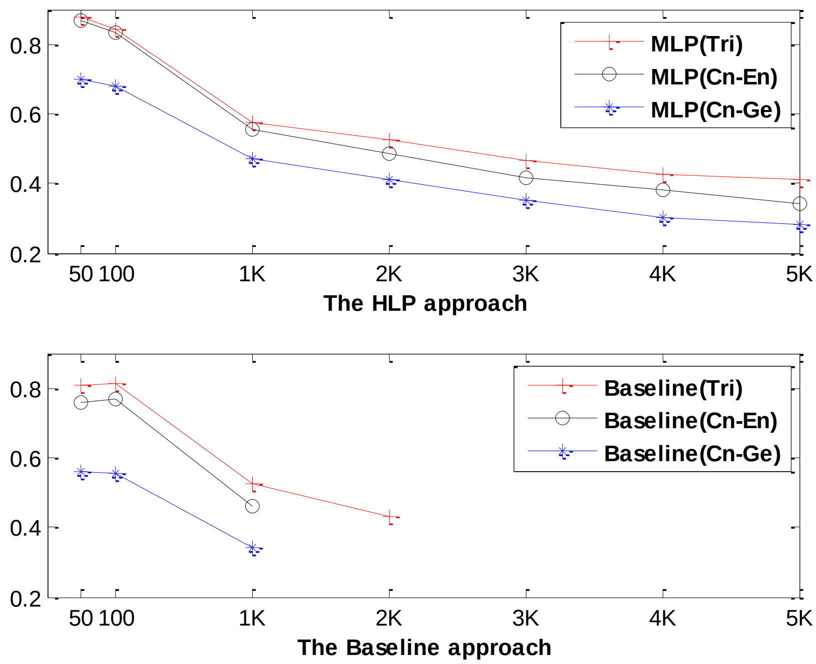


Figure 5.4 Evaluation on multilingual resources (precision)

As shown in Figure 5.4 and Figure 5.5, three findings are observed. First, since the quantity of English resources is much larger than that of German resources, the MLP approach gains more benefits from English resources. In terms of the precisions at the top-ranked Chinese words (i.e., top50, top100 and top1000), the MLP(Tri) approach slightly outperforms the MLP(Cn-En) approach. Compared with the MLP(Cn-En) and MLP(Cn-Ge) approaches, the precision improvement in the MLP(Tri) approach when incorporating multilingual resources is clearly observed for the low-ranked words (top1000~top5000). This indicates that the MLP approach is able to learn certain numbers of high-quality sentiment words based on bilingual resources. The contribution of multilingual resources to the MLP approach is to improve the coverage of the learned sentiment lexicons. Thus we observe the increase in recall as illustrated in Figure 5.5.

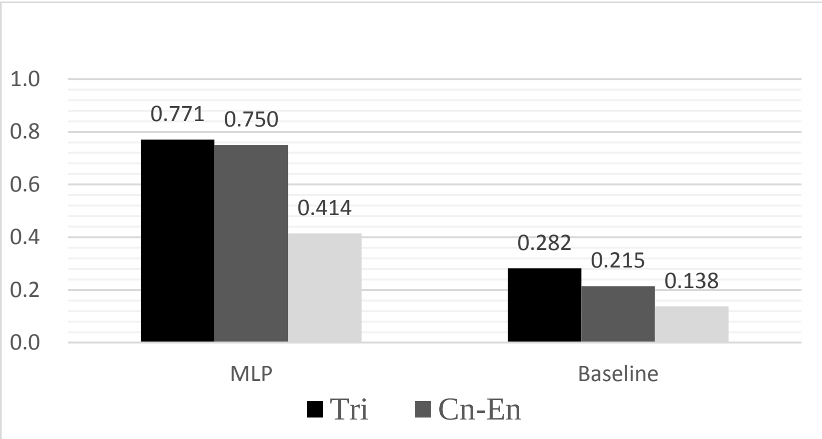


Figure 5.5 Evaluation on multilingual resources (recall)

Second, the MLP approach achieves a significant improvement in comparison with the baseline approach. With the MLP approach, the sentiment information gradually propagates over the multilingual graph. In the baseline approach, the sentiment information transfers through the connections once only.

Thus, the word sentiment information in the MLP approach is more refined, which leads to performance improvement. In Figure 5.5, the recall of the MLP approach is significantly higher than the recall of the Baseline approach. This means that the MLP approach learns much more sentiment words.

Third, when manually analyzing 100 sentiment words (the top-50 words from the positive and negative lists, respectively) learned from the baseline approach, it is found that some wrongly-predicted words are domain related words e.g., “决议” (provision, item, clause) and “实现” (implementation, achievement). Since the dataset is extracted from the official documents of the United Nations, these words occur frequently. These frequently occurred words occasionally connect to certain sentiment words, consequently they get the sentiment information and are predicted as sentiment words. This side effect is reduced in the MLP approach, where the sentiment information is extensively refined and is propagated to more words in the multilingual graph. When more words accurately obtain the sentiment information, the domain related words have less opportunities to get a high rank in the lists. Even so, some of the words still appear in the top-200 ranked lists in the MLP approach. It is the fundamental problem that the connections in the word graph may link the sentiment words to the non-sentiment words. Among them, some may be the noisy links, which wrongly link the sentiment words to the non-sentiment words. Some others may be the reasonable ones in the sense that they connect the words with similar meanings, but they are not suitable for sentiment lexicon learning, like the connection linking “blue (low in spirits)” and “蓝色 (the color blue)”. This problem exists so long as the non-sentiment-bearing connections exist in the multilingual graph. To further alleviate this side effect, the approach that can filter out the non-sentiment-bearing connections or can enhance the weights of the sentiment-bearing connections is expected.

5.4.2 Evaluation on Sentiment-oriented Label Propagation

This set of experiments is to evaluate the proposed SOLP approach. The proposed SOLP approach is compared with the following existing work.

RCL: [Ji et al. 2011] present the work on classifying the bibliographic data (including author, publications and venues e.g., conferences and journals) into research communities. The RCL approach they propose also gradually updates the graph connections. We apply the RCL approach to multilingual sentiment lexicon learning. The RCL approach assumes that the highly ranked words should play more important roles in the next round of propagation. The weight update is formulated as

$$w_{ip,jq}^{(t+1)} = w_{ip,jq}^{(t)} * (\sigma + \sqrt{\frac{P(x_{ip}|\mathrm{c}_k)}{\max_p\left(P(x_{ip}|\mathrm{c}_k)\right)} * \frac{P(x_{jq}|\mathrm{c}_k)}{\max_q\left(P(x_{jq}|\mathrm{c}_k)\right)}})$$

where $P(x_{ip}|c_k)$ denotes the ranking distribution of the word x_{ip} on the sentiment class c_k.

This set of experiments is based on the multilingual resources. As shown in Figure 5.6, compared with the MLP approach, the RCL and SOLP approaches archive better performances in precision. It verifies our assumption that the graph update benefits sentiment words learning. Actually, both RCL and SOLP approaches involve certain processes like relevance feedback, which has been claimed to be effective in information retrieval [Büttcher et al. 2010] and feature selection [Gao et al. 2011; Gao et al. 2012]. Specifically, the RCL approach performs a pseudo relevance feedback. It regards the words in top ranks as sentiment words and highlights the roles of the newly-learned confident sentiment words for further improvement. In contrast, the sentiment-oriented graph update in the SOLP approach is like an explicit relevance feedback. The SOLP approach leverages the sentiment seed words to automatically assess the sentiment-oriented consistency of the two connected words. In each iteration if the sentiment-oriented consistency of the two words connected by a connection is high, the weight of the connection increases. In the next iteration, this connection will become more important for label propagation and the connected words tend to obtain more sentiment information. In these two approaches, the confident sentiment score or the high sentiment-oriented consistency from one iteration is continuously leveraged in the learning processes of the next iteration, as if is in relevance feedback. Thus the superiorities of the RCL and SOLP approaches in sentiment lexicon learning is observed.

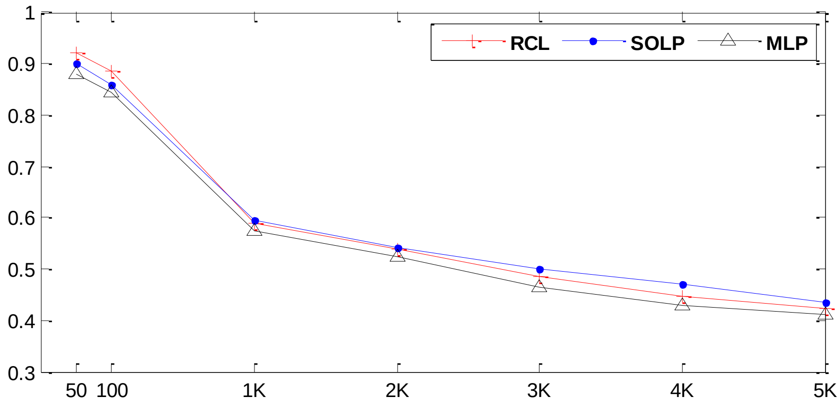


Figure 5.6 Evaluation on sentiment-oriented label propagation (precision)

Furthermore, the RCL approach outperforms the SOLP approach in the Top-K (K < 100) precision, while the performance of the SOLP approach is superior to that of the RCL approach when K>1000. In the RCL approach, when a word gets a high score, the weights of its connections tend to increase and this word will get an even higher score in the next iteration. It seems that the RCL approach focuses more on the higher-ranked nodes. It aims to ensure the top ranked words are the most confident sentiment words. In contrast, the SOLP approach updates the weights of the connections with the sentiment-oriented consistencies of the connected words. The graph update in the SOLP approach focuses more on word connections. It aims to ensure the sentiment information to be accurately propagated along the connections that link the two sentiment words. The sentiment-oriented graph update focuses on all the connections in the graph, rather than the top-ranked nodes in RCL. Thus, the RCL approach outperforms the SOLP approach in the top-ranked evaluation, while the SOLP approach is better on the entire rank lists. For sentiment lexicon learning, we not only require the relative-high precision, but also expect a proper coverage. Table 5.2 below shows the recall improvement of the SOLP approach in comparison with the RCL approach. It is more practical to use the proposed SOLP approach in multilingual sentiment lexicon learning, while the RCL approach seems more suitable for the applications that require high precisions at the top ranked items, like bibliographic information ranking [Ji et al. 2011]. In conclusion, the SOLP approach is responsible to automatically reweight the connections for sentiment lexicon learning. It is the advantage of the SOLP approach that enables us to focus on word connection mining, like employing the hypernym and hyponym relations in the learning process.

The previous approaches determine the sentiment polarities of the words according to the information in the word graphs. They are referred to as the graph-based approaches. In the following experiments, the model-based co-training approach presented in Chapter 4 is also implemented for comparison.

COR: Three classifiers are developed for English, German and Chinese. If a word in a language is confidently predicted on a sentiment polarity, this word is regarded as a new training word for its own language in the next iterations. Meanwhile, through word alignment, the aligned words in the other two languages are used to update the corresponding training datasets of these two languages. The sentiment seed words are used as the training data. After co-training, the classifiers generate multilingual sentiment lexicons by predicting the sentiment polarities of the unlabeled words in each language.

	RCL	SOLP	MLP	COR
Recall	0.775	0.802	0.771	0.804

Table 5.2 Evaluation on sentiment-oriented label propagation (recall)

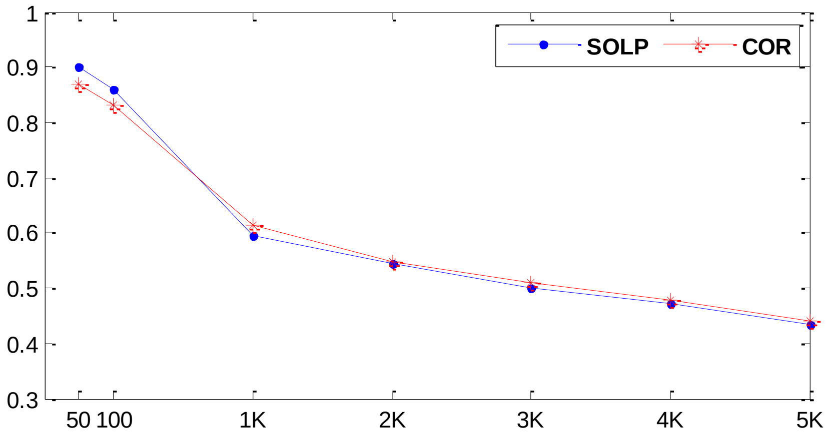


Figure 5.7 Comparison with the co-training approach (precision)

Figure 5.7 and Table 5.2 show that the co-training and SOLP approaches achieve comparable results in multilingual sentiment lexicon learning. In the co-training approach, though we do not explicitly reweight the graph connections, the classification algorithm (e.g., SVM) can automatically select the representative features for training. To a certain degree, this automatic feature selection process roughly corresponds to the process of automatically selecting sentiment-bearing connections for sentiment information transferring. Thus the two approaches are able to achieve comparable results and both are effective in multilingual sentiment lexicon learning.

Precision	P@ 50	P@ 100	P@ 1K	P@ 2K	P@ 3K	P@ 4K	P@ 5K
MLP	0.880	0.845	0.574	0.525	0.464	0.428	0.412
SOLP	0.900	0.860	0.596	0.543	0.501	0.472	0.436
SOLP(Inter)	0.900	0.860	0.584	0.538	0.494	0.465	0.424
SOLP(Intra)	0.890	0.850	0.576	0.528	0.484	0.433	0.414

Table 5.3 Influence of the intra/inter-language sub-graph update (precision)

	MLP	SOLP	SOLP(Inter)	SOLP(Intra)
Recall	0.771	0.802	0.792	0.778

Table 5.4 Influence of the intra/inter-language sub-graph update (recall)

5.4.3 Influence of the Intra- and Inter- Language Sub-graph Update

The multilingual graph in MSLL contains m intra-language sub-graphs and $m(m-1)$ inter-language sub-graphs. The SOLP approach updates all the intra/inter-language sub-graphs at the same time. This set of experiments is to examine the influence of the different sub-graph update. In the SOLP(Intra) approach, the intra-language sub-graphs are updated, while in the SOLP(Inter) approach the inter-language sub-graphs are updated.

As shown in Table 5.3 and Table 5.4 in the next page, the SOLP(Inter) approach achieves a comparable result with the SOLP approach, and it outperforms the SOLP(Intra) approach. In other words, the SOLP approach benefits more from the inter-language sub-graph update. In MSLL, the intra-language relations are built based on the synonym relations. These relations are elaborately compiled by human experts. The inter-language relations are built according to word alignments. These word alignments are automatically extracted from the parallel sentences. Though word alignments benefit to the recall as discussed in Chapter 3, they may include more noises in comparison with the human-created synonym relations. Thus, the update of the inter-language sub-graphs is able to make more contributions for multilingual sentiment lexicon learning.

Precision	P@ 50	P@ 100	P@ 1K	P@ 2K	P@ 3K	P@ 4K	P@ 5K
MLP	0.880	0.845	0.574	0.525	0.464	0.428	0.412
SOLP(Local)	0.890	0.850	0.577	0.525	0.467	0.428	0.414
SOLP	0.900	0.860	0.596	0.543	0.501	0.472	0.435
SOLP(All)	0.900	0.865	0.592	0.553	0.498	0.454	0.427

Table 5.5 Influence of the graph update strategies (precision)

5.4.4 Evaluation on Sub-graph Update Strategies

In the SOLP approach, for a sub-graph of one language (the intra-language sub-graph) or of two languages (the inter-language sub-graph), the sentiment seed words from all the other languages will be used to reweight the sub-graph in our approach. For example, the German sentiment seed words are used to update the inter-language sub-graph of English and Chinese. In this situation, the German sentiment seed words are named as the cross-lingual sentiment seed words for the inter-language sub-graph of English and Chinese. In the following set of experiments, different graph update strategies are compared. In the SOLP(Local) approach, for a sub-graph of one language (the intra-language sub-graph) or of two languages (the inter-language sub-graph), its update is done by the sentiment seed words from that/those language(s). These sentiment seed words are deemed local to the sub-graph. For example, the English sentiment seed words are the local sentiment seed words for the intra-language sub-graph of English. The English and German sentiment seed words are the local sentiment seed words for the inter-language sub-graph of English and German. The SOLP(All) approach updates each sub-graph with the sentiment seed words from all the languages.

	MLP	SOLP(Local)	SOLP	SOLP(ALL)
Recall	0.771	0.773	0.802	0.795

Table 5.6 Influence of the graph update strategies (recall)

As for the MLP approach, the performance of the SOLP(Local) approach is not improved by updating the sub-graphs with local sentiment seed words as shown in Table 5.5 and Table 5.6. As for the SOLP approach, the performance of the SOLP(All) approach slightly declines in both precision (decline 0.3% in average) and recall (decline 0.7% in average). The main difference between these two update strategies lies in whether the local sentiment seed words are used in the sub-graph update or not. For a sub-graph, in average the initial similarities of the unlabeled words to the local sentiment seed words are much sparser than those to the cross-lingual sentiment seed words. The similarities to the local sentiment seed words are initialized by the word synonym information, and the similarities to the cross-lingual sentiment seed words are initialized by the word alignment information. As illustrated in Table 5.7, initially for a word, it connects to 0.430 local sentiment seed words, but connects to 2.447 cross-lingual sentiment seed words in average. The similarities to the local sentiment seed words are sparser. Thus the influence of the local sentiment seed words is less than that of the cross-lingual sentiment seed words.

Language	Cross-lingual	Local
English word	2.370	0.645
Chinese word	2.200	0.325
German word	2.770	0.320
Average	2.447	0.430

Table 5.7 Average numbers of the initially-connected local and cross-lingual sentiment seed words

This set of experiments confirms the effectiveness of the cross-lingual sentiment seed words in the sentiment-oriented graph update. To perform this cross-lingual update, at least a third language is needed to update the inter-language sub-graph. To a certain degree, this claims the necessary of incorporating multiple language resources (at least three languages) in the proposed SOLP approach.

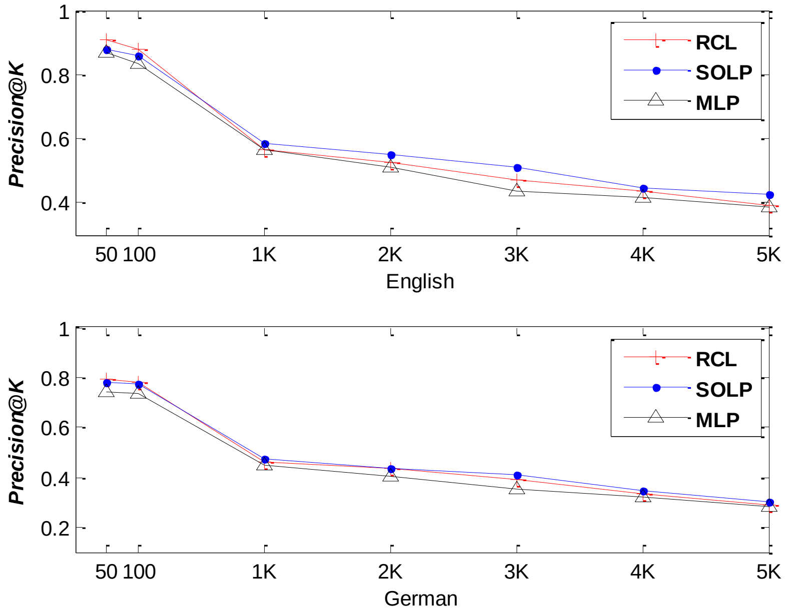


Figure 5.8 Evaluation on learned English and German sentiment lexicons (precision)

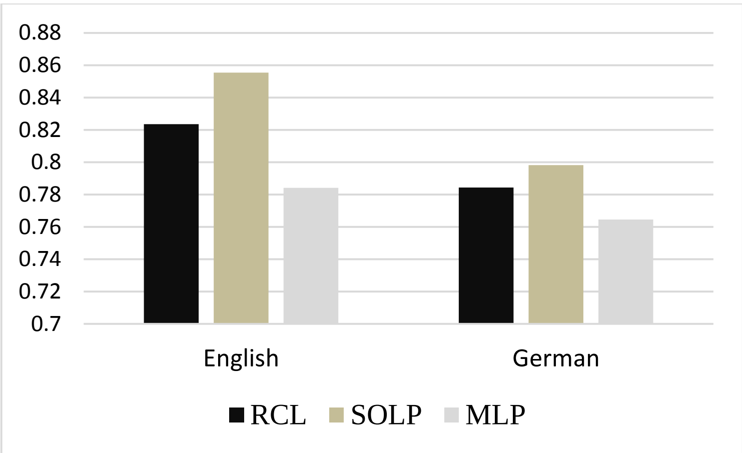


Figure 5.9 Evaluation on learned English and German sentiment lexicons (recall)

5.4.5 Evaluation on Learned English and German Sentiment Lexicons

MSLL can learn the sentiment lexicons for multiple languages at the same time. Similar to the evaluation on the learned Chinese sentiment lexicon, the learned English and German sentiment lexicons are also evaluated in terms of Precision@K and Recall. The performances of the MLP, SOLP and RCL approaches are shown in Figure 5.8 and Figure 5.9. The trends of these curves on English and German sentiment lexicon evaluation are similar to that on Chinese sentiment lexicon evaluation in Section 5.4.2. The SOLP and RCL approaches, in comparison with the MLP approach, achieve better performances with the help of the sentiment-oriented graph update. Furthermore, the RCL approach slightly outperforms the SOLP approach in the top-K (K<500) words, while the SOLP approach shows a better performance in the entire ranked lists. Besides, the performance of the learned English sentiment lexicon is significantly superior to that of German. In these approaches, the quantity of the English resources is much larger than that of the German ones. Thus, the performance of the English sentiment lexicon is better.

NTCIR	English	Chinese
Positive	937	1,218
Negative	1,923	944
Total	2,860	2,162

Table 5.8 Detail about the NTCIR dataset

5.5 Further Discussion

By now, I have introduced all my work on cross-lingual sentiment lexicon learning (CSLL). The relations between these three pieces of work are as following.

In Bilingual graph based Sentiment Lexicon Learning (BSLL), we address the two main issues in CSLL, *cross-lingual sentiment information mapping* and *word sentiment representation*. Word alignment is leveraged to build the inter-language relation. The sentiment polarity of a word is determined by the sentiment polarities of all its connected words in bilingual word graph. In Co-training based Bilingual Sentiment Lexicon Learning (CBSLL), we improve word sentiment representation in BSLL. More information (e.g., word definition, structure balance) is explored to determine word sentiment polarity. Multilingual based Sentiment Lexicon Learning (MSLL) is the generation of BSLL, from bilingual word graph to multilingual word graph. Furthermore, we improve sentiment information mapping in BSLL. Sentiment-Oriented Label Propagation (SOLP) is proposed to automatically increase the weights of sentiment-bearing connections for more accurate sentiment information mapping.

These three pieces of work provide a systematic of studies on CSLL, which fills in the insufficiency of the study on CSLL, to a great extent. Meanwhile, the proposed new models and the extended framework also contribute to the related research topics, e.g., the prediction task in some online social networks (or media), which involve the signed graphs.

5.6 Sentiment classification with learned sentiment lexicon

In all the above work, the learned sentiment lexicons are evaluated with Precision and Recall. Now, it is the time to apply the learned sentiment lexicons to a practical application to further examine the qualities of the learned sentiment lexicons. Sentiment classification is one of the most extensively studied topics in the communities of sentiment analysis [Pang and Lee 2008]. In this section, the words in the learned sentiment lexicons are used as the features in sentence level sentiment classification.

Dataset: The NTCIR labeled news articles are used for sentiment classification. They contain the sentiment labeled sentences in both Chinese and English [Seki et al. 2008; Seki et al. 2009]. The sentences that are with positive or negative labels are extracted for evaluation. The numbers of extracted sentences are shown in Table 5.8.

The sentiment words identified by the SBLP approach presented in Chapter 3, by the co-training approach presented in Chapter 4 and by the SOLP approach presented in this chapter are used as the features in sentiment classification. They are compared with the following two baselines.

Baseline_Cor: In this approach, the words including the sentiment words and the non-sentiment words in the NTCIR dataset are extracted. They are ranked according to their appearance frequencies. The top-*N* words are used as the classification features. As reported in [Pang et al. 2002], unigram features are more effective than bigram features for English sentiment classification. Thus, this baseline approach employs unigrams as English classification features. The experimental results in [Li et al. 2006; Li and Sun 2007] show that Chinese bigrams also convey the informative semantic information and bigram features perform well in Chinese text classification and Chinese sentiment classification. Therefore, this approach employs both unigrams and bigrams as features for Chinese sentiment classification.

Baseline_Lex: In this approach, the words in existing sentiment lexicons are used as features. For Chinese, the sentiment words in *HowNet* are used as the features. From HowNet, I collect 836 positive words and 1,254 negative words. For English, the sentiment words from *sentiWordNet*, are used as the features. From sentiWordNet, 4,979 positive and 8,421 negative words are collected. The sentiment words are ranked according to their sentiment scores. The top-*N* sentiment words are used as the classification features.

The *SMO classifier* [Platt 1999] in *Weka*[35] is applied to perform 10-fold cross-validation on the NTCIR labeled sentences. The accuracy is shown in Figure 5.10. For Chinese sentiment classification, the proposed SBLP, co-training and SOLP approaches all achieve very promising performances though the features and sentences are from different corpora. This demonstrates that the generated sentiment lexicons are adaptive and the qualities of them are good enough for Chinese sentiment classification. From English sentiment classification, the SBLP, co-training and SOLP approaches are superior to the Baseline_Cor approach. The performances of the SOLP and co-training approaches are comparative to the performance of the Baseline_Lex approach after using about the top-5K sentiment words. The reason is that sentiWordNet already covers amounts of sentiment words. I need to involve more learned words to compare with its coverage.

According to the evaluation results provided in Section 5.4.2 and Section 4.4.1, the SOLP and co-training approaches outperform the SBLP approach in sentiment lexicon learning. In the evaluation of this section, the SOLP and co-training approaches also achieve the consistent improvement compared with the SBLP approach in sentiment classification. This not only reclaims the

---

[35] www.cs.waikato.ac.nz/ml/weka

superiority of the SOLP and co-training approaches to the SBLP approach, but also suggests that if an approach has a good performance in sentiment lexicon learning, the sentiment lexicon generated by this approach tends to perform well in sentiment classification.

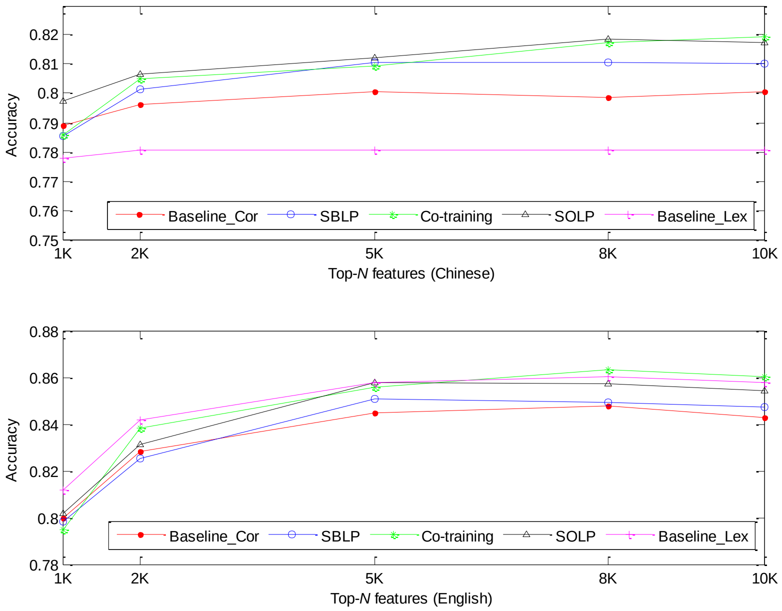


Figure 5.10 Evaluation on sentence-level sentiment classification

5.7 Chapter Summary

In this chapter, the work on multilingual sentiment lexicon learning is presented. A multilingual graph is established by incorporating the resources from multiple languages. The sentiment-oriented label propagation approach is proposed to transfer sentiment information among languages and to update the original word semantic-oriented connections to the sentiment-oriented connections. Extensive experiments are conducted on English, Chinese and German sentiment lexicon learning. The results demonstrate the effectiveness of the SOLP approach. The major contribution of this work is that the idea behind the SOLP approach can also be applied to the other problems where the specific connections are lack or costly to acquire. The proposed SOLP approach can automatically update the connections for the specific use of the problems. The learned sentiment lexicons are finally evaluated on sentence level sentiment classification. For Chinese and English sentiment classification, the approaches based on the learned sentiment lexicons perform well in classification, especially in Chinese sentiment classification. This indicates the advantages of the learned sentient lexicons.

Chapter 6 Conclusion and Future Work

In this chapter, I wrap up the dissertation by assembling the three main chapters into a complete account of cross-lingual sentiment lexicon learning and indicate the future directions of my work.

6.1 Research Summary

In this dissertation, a systematic study is conducted for cross-lingual sentiment lexicon learning. First of all, the important role of sentiment lexicons in the researches of sentiment analysis is discussed. Our technology development for cross-lingual sentiment lexicon learning is then presented.

Our work starts with bilingual graph based sentiment lexicon learning, which learns the sentiment lexicon for a target language by taking advantage of widely abundant and available sentiment resources in a source language. In this work, the label propagation based approach is developed to induce the sentiment lexicon for the target language. Experimental results indicate that the sentiment information can be extensively propagated and precisely refined and the proposed approach shows its advantages in both precision and recall.

To combine more word information in sentiment lexicon learning, we work on co-training based bilingual sentiment lexicon learning. For a word, aside from the text-related features explored from its syntactic and semantic information, the graph-related features are also explored from its graph structure in the word graph. Unlike the static text-related features, which do not change in the learning processes, the graph-related features are allowed to be updated when the word graph accumulates more and more sentiment information. With these features, the co-training based approach is proposed to generate and/or expand the bilingual sentiment lexicons for two languages simultaneously with the available sentiment information in both languages. Experimental results indicate the effectiveness of the text-related and graph-related features in bilingual sentiment lexicon learning under the co-training framework.

To propagate the sentiment information more accurately, the work on multilingual sentiment lexicon learning is carried out. In this work, multilingual resources (including sentiment resources), on the one hand, are incorporated for their abundant sentiment information, and on the other hand, are used to guide the sentiment-oriented graph update. The proposed sentiment-oriented label propagation approach propagates the sentiment information between multiple languages and automatically update the semantic-oriented connections to the sentiment-oriented connections.

The learned sentiment lexicons are then evaluated on an application-oriented sentiment classification task. The experiments conducted on the NTCIR dataset have demonstrated the effectiveness of the learned sentiment lexicons in sentence level sentiment classification.

6.2 Technical Highlights

The following list itemizes the major highlights of our major work on cross-lingual sentiment lexicon learning.

1. Bilingual graph based sentiment lexicon learning
    1) A signed bilingual graph is built on the words in the two languages with their positive relations (i.e., the word alignment and synonym relations) and negative relations (i.e., antonym relations). Based on the signed bilingual graph, the signed bilingual label propagation (SBLP) approach is able to propagate the sentiment information along both positive and negative relations. Worth mentioning, the SBLP approach can also be applied to the researches in online social networks that involve signed graphs, like "Trust vs. Distrust" prediction [Guha et al. 2004; DuBois et al. 2011], "positive vs. negative" link prediction [Leskovec et al. 2010] and opinion influence [Li et al. 2012].
    2) The word alignment information derived from parallel sentences is leveraged to set up the inter-language relations. Through word alignment, more inter-language relations are built, which is proven to benefit the sentiment information transferring between languages. The word alignment information provides a new way to bridge the gap of the two languages. It can be leveraged in the other cross-lingual researches, like cross-lingual sentiment classification.
2. Co-training based bilingual sentiment lexicon learning
    1) To the best of our knowledge, we first incorporate the graph-related features in sentiment lexicon learning. Apart from the text-related features (e.g., the WordNet based features), the graph characteristics are

explored as the graph-related features (e.g., the structure balance features). The proposed co-training approach allows the graph-related features to be updated with the newly deducted sentiment information from the unlabeled words. This update strategy makes the graph-related features gradually take more effect in word sentiment classification.

2) In the co-training approach, the learning processes of the two languages are associated by word alignment. In each iteration of learning, if an unlabeled word in one language is confidently predicted with a certain sentiment polarity, its aligned word(s) in the other language are also assigned the same sentiment polarity. These words with their sentiment polarities are then used to update the training datasets of the corresponding languages. In this way, the learned sentiment information from either one language is interchanged to the other one and the two learning processes are associated.

3. Multilingual sentiment lexicon learning

1) In this work, we propose and approach the issue of multilingual sentiment lexicon learning (MSLL) for the first time. Multilingual resources are employed and uniformly formalized with a multilingual word graph. The advantage of MSLL is that it utilizes the sentiment information from multiple languages in the learning process of any one language. As demonstrated in the evaluation, the multilingual resources significantly improve the recall of the learned sentiment lexicons.

2) A sentiment-oriented label propagation (SOLP) approach is developed to propagate the sentiment information between multiple languages. The significance of the SOLP approach is that it is able to automatically

reweight the initial semantic-oriented connections to the sentiment-oriented connections. It leverages the benchmark sentiment words to measure the sentiment-oriented consistencies of the connections, and then uses these sentiment-oriented consistencies to update the weights of the connections. This idea can also be applied to the other problems where the specific connections are lack or costly to acquire.

6.3 Future Directions

The current work can be further extended along several directions. Besides developing more sophisticated approaches, two major directions are listed in the following.

One direction is to carry out cross-lingual sentiment lexicon learning for the specific languages (e.g., for Chinese and/or German). The proposed approaches in this dissertation are language-independent. We do not consider the language characteristics. In the future, interested researchers can further extend our work to learn sentiment lexicons for the specific languages. They can refine or expand the learned sentiment lexicons with more pragmatics knowledge from the specific languages. For instance, the four-character phrase (e.g., “心想事成” (all wishes come true) and “尽善尽美” (reach the perfection)) is an important component in the Chinese vocabulary. A large percentage of the four-character phrases contain certain sentiment polarities. Considering this language characteristics, researchers may like to incorporate more additional resources to further refine the sentiment polarities of the four-character phrases. Another direction is to associate cross-lingual sentiment lexicon learning with the other researches (e.g., sentiment extraction) in a unified framework. For example, as sentiment aspect words and sentiment words usually appear at the same time in most opinionated text units, we may want to extract sentiment aspect words simultaneously with sentiment words. Given Chinese opinioned text units, we can translate them into English. Then, existing approaches in English sentiment words and sentiment aspect words identification are able to use the translated texts. With these identified English sentiment (aspect) words and certain bilingual information, we can extract the Chinese sentiment words and Chinese sentiment aspect words in the original texts simultaneously.

At the end of this dissertation, I would like to say that in comparison with the other kinds of lexicons, the sentiment lexicon evolves much faster. Researchers will constantly confront new challenges, e.g., different sentiment expressions in the cyber texts. In the future, I will make continued effort to promote the development of sentiment lexicon learning.

References

Abbasi, A., Chen, H., and Salem, A. 2008. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems*, 26(3):1–34.

Aggarwal, C. C., and Zhai, C. 2012. A survey of text classification algorithms. Springer US.

Aisopos, F., Papadakis, G., Tserpes, K., and Varvarigou, T. 2012. Textual and contextual patterns for sentiment analysis over microblogs. *In Proceedings of the 21st International Conference Companion on World Wide Web*, pages 453–454. Lyon, France.

Alpaydin, E. 2010. Introduction to Machine Learning (the second edition). *The MIT Press*.

Amine, B. M., and Mimoun, M. 2007. WordNet based Multilingual Text Categorization. *Journal of Computer Science*, 6(4):52–59.

Anbananthen, K. S. M., and Elyasir, A. M. H. 2013. Evolution of Opinion Mining. *Australian Journal of Basic and Applied Sciences*, 7(6):359–370.

Baccianella, S., Esuli, A., and Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *In Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2200–2204. Malta, Malta.

Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., and Varma, V. 2012. Mining sentiments from Tweets. *In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 11–18. Jeju, Republic of Korea.

Banea, C., Mihalcea, R., and Wiebe, J. 2010. Multilingual subjectivity: are more languages better? *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 28–36. Beijing, China.

Bickerstaffe, A., and Zukerman, I. 2010. A Hierarchical Classifier Applied to Multi-way Sentiment Detection. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 62–70. Beijing, China.

Blair-Goldensohn, S. 2008. Building a sentiment summarizer for local service reviews. *In WWW Workshop on NLP in the Information Explosion Era*. Beijing China.

Blei, D. M. 2012. Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blum, A., and Mitchell, T. 1998. Combining Labeled and Unlabeled Data with Co-training. *In Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100. Madison, Wisconsin, USA.

Bollen, J., Mao, H., and Zeng, X. 2007. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Branavan, S. R. K., Chen, H., Eisenstein, J., and Barzilay, R. 2008. Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research*, 34(1):569–603.

Brody, S., and Elhadad, N. 2010. An Unsupervised Aspect-sentiment Model for Online Reviews. *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for*

Computational Linguistics, pages 804–812. Los Angeles, California, USA.

Brooke, J., Tofiloski, M., and Taboada, M. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 50–54. Borovets, Bulgaria.

Bütcher, S., Clarke, C. L. A., and Cormack, G. V. 2010. Information Retrieval: Implementing and Evaluating Search Engines. *Cambridge, Mass.: MIT Press*.

Campsvalls, G., Bandos, T. V., and Zhou, D. 2007. Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 45:2044–3054.

Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. 2007. Language resources and linguistic theory: Typology, second language acquisition, English linguistics*: Franco Angeli Editore*.

Chapelle, O., Scholkopf, B., and Zien, A. 2006. Semi-Supervised Learning. *Cambridge, Mass.: MIT Press*.

Chen, B., Zhu, L., Kifer, D., and Lee, D. 2010. What is an Opinion About? Exploring Political Standpoints using Opinion Scoring Model. *In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. Atlanta, Georgia, USA.

Chen, Y., and Eisele, A. 2012. MultiUN v2: UN Documents with Multilingual Alignments. *In Proceedings of the eighth international conference on Language Resources and Evaluation*, pages 2500–2504. Istanbul, Turkey.

Cheung, J. C. 2008. Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection. University of

British Columbia.

Choi, Y., and Cardie, C. 2010. Hierarchical Sequential Learning for Extracting Opinions and Their Attributes. *In Proceedings of the ACL 2010 Conference*, pages 269–274. Uppsala, Sweden.

Chomsky, N. 2002. Syntactic Structures (the second edition). *Walter de Gruyter*.

Cortes, C., and Vapnik, V. 1995. Support-Vector Networks. *Journal Machine Learning*, 20(3):273–297.

Das, D., and Martins, A. F. T. 2007. A Survey on Automatic Text Summarization. Carnegie Mellon University.

Das, D., and Petrov, S. 2011. Unsupervised Part-of-speech Tagging with Bilingual Graph-based Projections. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609. Portland, Oregon, USA.

Dave, K., Lawrence, S., and Pennock, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th International Conference on World Wide Web*, pages 519–528. Budapest, Hungary.

Davidov, D., Tsur, O., and Rappoport, A. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. *In Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Beijing, China.

Ding, X., Liu, B., and Yu, P. S. 2008. A Holistic Lexicon-based Approach to Opinion Mining. *In Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240. Palo Alto, California, USA.

Dong, Z., Dong, Q., and Hao, C. 2010. HowNet and Its Computation of Meaning.

In Proceedings of the 23rd International Conference on Computational Linguistics, pages 53–56. Beijing, China.

DuBois, T., Golbeck, J., and Srinivasan, A. 2011. Predicting Trust and Distrust in Social Networks. *In Proceedings of the IEEE 3rd International Conference on Social Computing*, pages 418–424. Beijing, China.

Duh, K., Fujino, A., and Nagata, M. 2011. Is Machine Translation Ripe for Cross-lingual Sentiment Classification? *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 429–433. Portland, Oregon, USA.

Easley, D., and Kleinberg, J. 2010. Positive and negative relationships Networks, Crowds, and Markets: Reasoning about a Highly Connected World, pages 119–152. Cambridge University Press.

Esuli, A., and Sebastiani, F. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *In Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 417–422. Genoa, Italy.

Esuli, A., and Sebastiani, F. 2007. Random-Walk Models of Term Semantics: An Application to Opinion-Related Properties. *In Proceedings of the 3rd Language and Technology Conference*, pages 221–225. Poznan, Poland.

Fano, R. M. 1961. Transmission of Information: A Statistical Theory of Communications. *Cambridge, Mass., USA: MIT Press*.

Feldman, R., Rosenfeld, B., Barhaim, R., Fresko, M., Shapiro, D. G., and Fromherz, M. P. J. 2011. The Stock Sonar - Sentiment Analysis of Stocks Based on a Hybrid Approach. *In Proceedings of the 23rd Conference on Innovative Applications of Artificial Intelligence*, pages 1642–1647. San

Francisco, California, USA.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 2:1289–1305.

Ganesan, K., Zhai, C., and Han, J. 2010. Opinosis: A Graph-based Approach to Abstractive Summarization of Highly Redundant Opinions. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348. Beijing, China.

Gao, D. 2012. Opinion Influence and Diffusion in Social Network. *In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 997–997. Portland, Oregon, USA.

Gao, D., Li, W., Ouyang, Y., and Zhang, R. 2012. LDA-Based Topic Formation and Topic-Sentence Reinforcement for Graph-Based Multi-document Summarization. *In Proceeding of the 8th Asia Information Retrieval Societies Conference*, pages 376–385. Tianjin, China.

Gao, D., Li, W., and Zhang, R. 2012. Efficient Feedback-based Feature Learning for Blog Distillation as a Terabyte Challenge. *In Proceeding of the 24th International Conference on Computational Linguistics*, pages 147–154. Bombay, Indian.

Gao, D., Zhang, R., Li, W., Lau, R. Y. K., and Wong, K. F. 2011. Learning features through feedback for blog distillation. *In Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1085–1086. Beijing, China.

Ghahramani, Z., and Jordan, M. I. 1997. Factorial Hidden Markov Models. *Machine Learning*, 29(2-3):245–273.

Griffiths, T. L., and Steyvers, M. 2003. Prediction and Semantic Association. *Advances in Neural Information Processing Systems*, 15:(11–18).

Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. 2004. Propagation of Trust and Distrust. *In Proceedings of the 13th International Conference on World Wide Web*, pages 403–412. New York, NY, USA.

Gupta, V., and Lehal, G. S. 2010. A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3):258–268.

Hassan, A., Abu-Jbara, A., Jha, R., and Radev, D. 2011. Identifying the semantic orientation of foreign words. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 592–597. Portland, Oregon, USA.

Hatzivassiloglou, V., and McKeown, K. R. 1998. Predicting the semantic orientation of adjectives. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181. Madrid, Spain.

Heider, F. 1946. Attitudes and cognitive organization. *Journal of Psychology*, 21:107–112.

Hobbs, J. R., and Riloff, E. 2010. Information Extraction (the second edition). *In Handbook of Natural Language Processing*. CRC Press, Taylor and Francis Group.

Hofmann, T. 1999. Probabilistic Latent Semantic Indexing. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. Berkeley, California, USA.

Hsieh, C. Y., Chiang, K. Y., and Dhillon, I. S. 2012. Low rank modeling of signed networks. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 507–515. Beijing, China.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. *In Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. Seattle, WA, USA.

Hu, Y., and Li, W. 2011. Document sentiment classification by exploring description model of topical terms. *Computer Speech and Language*, 25(2):386–403.

Ji, M., Han, J., and Danilevsky, M. 2011. Ranking-Based Classification of Heterogeneous Information Networks. *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining Pages*, pages 1298–1306. San Diego, California, USA.

Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. 2010. Graph Regularized Transductive Classification on Heterogeneous Information Networks. *In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Barcelona, Spain.

Jiang, J. Q. 2011. Learning Protein Functions from Bi-relational Graph of Proteins and Function Annotations. *Algorithms in Bioinformatics*, 6833:128–138.

Jiang, J. Q., and McQuay, L. J. 2012. Predicting Protein Function by Multi-Label Correlated Semi-Supervised Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1059–1069.

Jin, W., and Ho, H. H. 2009. A Novel Lexicalized HMM-based Learning Framework for Web Opinion mining. *In Proceedings of the 26th Annual International Conference on Machine Learning*, pages 465–472. Montreal, Quebec, Canada.

Jo, Y., and Oh, A. H. 2011. Aspect and Sentiment Unification Model for Online Review Analysis. *In Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 815–824. Hong Kong, China.

Joachims, T. 1999. Making large-scale support vector machine learning practical. Advances in kernel methods, pages 169–184. MIT Press Cambridge.

Joachims, T. 2002. Learning to Classify Text Using Support Vector Machines. *Norwell, MA, USA: Kluwer Academic Publishers*.

Ju, S., Li, S., Su, Y., Zhou, G., Hong, Y., and Li, X. 2012. Dual word and document seed selection for semi-supervised sentiment classification. *In Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2295–2298. Maui, Hawaii, USA.

Kamps, J., Marx, M., Mokken, R. J., and Rijke, M. de. 2004. Using WordNet to Measure Semantic Orientations of Adjectives. *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1115–1118. Lisbon, Portugal.

Kennedy, A., and Inkpen, D. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125.

Kim, H. D., and Zhai, C. 2009. Generating Comparative Summaries of Contradictory Opinions in Text. *In Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 385–394.

Hong Kong, China.

Kim, S. M., and Hovy, E. 2004. Determining the Sentiment of Opinions. *In Proceedings of the 20th international conference on Computational Linguistics*, pages 355–363. Geneva, Switzerland.

Kim, S. M., and Hovy, E. 2006. Automatic identification of pro and con reasons in online reviews. *In Proceedings of the COLING/ACL on Main Conference Poster Sessions*, pages 483–490. Sydney, Australia.

Kobayashi, N., Iida, R., Inui, K., and Matsumoto, Y. 2006. Opinion Mining on the Web by Extracting Subject-Attribute-Value Relations. *In Proceedings of AAAI Spring Sympoia on Computational Approaches to Analyzing Weblogs*. Palo Alto, California, USA.

Kobayashi, N., Inui, K., and Matsumoto, Y. 2007. Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1065–1074. Prague, Czech Republic.

Kovelamudi, S., and Ramalingam, S. 2011. Domain Independent Model for Product Attribute Extraction from User Reviews using Wikipedia. *In Proceeding of the 5th International Joint Conference on Natural Language Processing*, pages 1408–1412. Chiang Mai, Thailand.

Ku, L. W., Lee, L. Y., and Chen, H. H. 2006. Opinion extraction, summarization and tracking in news and blog corpora. *In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*. Palo Alto, California, USA.

Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., De, E. W., and Albayrak, L.

S. 2010. Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization. *In Proceedings of the SIAM International Conference on Data Mining*, pages 559–570. Columbus, Ohio, USA.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Williamstown, MA, USA.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. 2010a. Signed Networks in Social Media. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370. Atlanta, Georgia, USA.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. 2010b. Predicting Positive and Negative Links in Online Social Networks. *In Proceedings of the 19th International Conference on World Wide Web*, pages 641–650. Raleigh, North Carolina, USA.

Li, B., Chen, J., Chen, X., Tang, X., and Qu, W. 2011. Sentimental Direction Analysis: A Framework for Chinese Sentiment Computation and Resource Construction. *International Journal of Knowledge and Language Processing*, pages 19–39.

Li, D., Shuai, X., Sun, G., Tang, J., Ding, Y., and Luo, Z. 2012. Mining Topic-level Opinion Influence in Microblog. *In Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1562–1566. Maui, Hawaii, USA.

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y. J., Zhang, S., and Yu, H. 2010. Structure-aware review mining and summarization. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages

653–661. Beijing, China.

Li, Jingyang, Sun, M., and Zhang, X. 2006. A Comparison and Semi-quantitative Analysis of Words and Character-bigrams As Features in Chinese Text Categorization. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 545–552. Sydney, Australia.

Li, Jun, and Sun, M. 2007. Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques. *In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, pages 393–400. Beijing, China.

Li, Shen, Graca, J. V., and Taskar, B. 2012. Wiki-ly Supervised Part-of-speech Tagging. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389–1398. Jeju Island, Korea.

Li, Shoushan, Lee, S. Y. M., Chen, Y., Huang, C.-R., and Zhou, G. 2010. Sentiment classification and polarity shifting. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 635–643. Beijing, China.

Liang, P., Taskar, B., and Klein, D. 2006. Alignment by Agreement. *In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111. Sydney, Australia.

Liu, B. 2012. Sentiment Analysis and Opinion Mining. *Morgan and Claypool Publishers*.

Liu, J., Cao, Y., Lin, C. Y., Huang, Y., and Zhou, M. 2007. Low-Quality Product

Review Detection in Opinion Summarization. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 334–342. Prague, Czech Republic.

Lu, B., Tan, C., Cardie, C., and Tsou, B. K. 2011. Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 320–330. Portland, Oregon, USA.

Lu, Y., and Zhai, C. 2008. Opinion Integration Through Semi-supervised Topic Modeling. *In Proceedings of the 17th International Conference on World Wide Web*, pages 121–130. Beijing, China.

Ma, H., Lyu, M. R., and King, I. 2009. Learning to Recommend with Trust and Distrust Relationships. *In Proceedings of the 3rd ACM Conference on Recommender Systems*, pages 189–196. New York, NY, USA.

Manevitz, L. M., and Yousef, M. 2002. One-class Svms for Document Classification. *Journal of Machine Learning Research*, 2:139–154.

Mcdonald, R., Hannan, K., Neylon, T., and Mike Wells, J. R. 2007. Structured Models for Fine-to-Coarse Sentiment Analysis. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. 2007. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. *In Proceedings of the 16th international conference on World Wide Web*, pages 171–180. Banff, Alberta, Canada.

Melville, P., Gryc, W., and Lawrence, R. D. 2009. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284. Paris, France.

Meng, X., Wei, F., Liu, X., Zhou, M., Xu, G., and Wang, H. 2012. Cross-lingual Mixture Model for Sentiment Classification. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 572–581. Jeju Island, Korea.

Meng, X., Wei, F., Xu, G., Zhang, L., Liu, X., Zhou, M., and Wang, H. 2012. Lost in Translations? Building Sentiment Lexicons using Context Based Machine Translation. *In Proceeding of the 24th International Conference on Computational Linguistics*, pages 829–838. Mumbai, India.

Mihalcea, R. 2004. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. *In Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*. Barcelona, Spain.

Mihalcea, R., Banea, C., and Wiebe, J. 2007. Learning Multilingual Subjective Language via Cross-Lingual Projections. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983. Prague, Czech Republic.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Moghaddam, S., and Ester, M. 2010. Opinion Digger: An Unsupervised Opinion Miner from Unstructured Product Reviews. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1825–1828. Toronto, ON, Canada.

Mooney, R. J., and Bunescu, R. 2005. Mining knowledge from text using information extraction. *ACM SIGKDD - Natural language processing and text mining*, pages 3–10. Baltimore, Maryland, USA.

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

Munteanu, D. S., and Marcu, D. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Na, J. C., Sui, H., Khoo, C., Chan, S., and Zhou, Y. 2004. Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews. *In Proceeding of the International Society for Knowledge Organization ISKO*, pages 49–54. London, England.

Nasukawa, T., and Yi, J. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *In Proceedings of the 2nd International Conference on Knowledge Capture*, pages 70–77. Sanibel Island, FL, USA.

Nenkova, A., and McKeown, K. 2012. A Survey of Text Summarization Techniques. *Mining Text Data*, pages 43–76. Springer US.

Nigam, K. 1999. Using maximum entropy for text classification. *In Proceeding of the 16th International Joint Conference on Artificial Intelligence, Workshop: Machine Learning for Information Filtering*, pages 61–67. Stockholm, Sweden.

Nishikawa, H., Hasegawa, T., Matsuo, Y., and Kikui, G. 2010a. Optimizing Informativeness and Readability for Sentiment Summarization. *In Proceedings of the ACL 2010 Conference Short Papers*, pages 325–330.

Uppsala, Sweden.

Nishikawa, H., Hasegawa, T., Matsuo, Y., and Kikui, G. 2010b. Opinion Summarization with Integer Linear Programming Formulation for Sentence Extraction and Ordering. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 910–918. Beijing, China.

Otterbacher, J., Erkan, G., and Radev, D. R. 2005. Using Random Walks for Question-focused Sentence Retrieval. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 915–922. Vancouver, British Columbia, Canada.

Ounis, I., Macdonald, C., and Soboroff, I. 2010. Overview of the TREC2010 Blog Track.

Pang, B., and Lee, L. 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*. Barcelona, Spain.

Pang, B., and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. Hanover, MA, USA.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86. Philadelphia, PA, USA.

Paul, M. J., Zhai, C., and Girju, R. 2010. Summarizing Contrastive Viewpoints in Opinionated Text. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76. Cambridge, Massachusetts, USA.

Platt, J. C. 1999. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. Cambridge, MA, USA: MIT Press.

Popescu, A. M., and Etzioni, O. 2005. Extracting product features and opinions from reviews. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. Vancouver, British Columbia, Canada.

Qiu, G., Liu, B., Bu, J., and Chen, C. 2009. Expanding Domain Sentiment Lexicon Through Double Propagation. *In Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1199–1204. Pasadena, California, USA.

Qiu, G., Liu, B., Bu, J., and Chen, C. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistic*, 37(1):9–27.

Quattoni, A., Wang, S., Morency, L. P., Collins, M., and Darrell, T. 2007. Hidden Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852.

Rabiner, L. R. 1990. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2): 267–296.

Rao, D., and Ravichandran, D. 2009. Semi-supervised polarity lexicon induction. *In Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 675–682. Athens, Greece.

Remus, R., Quasthoff, U., and Heyer, G. 2010. SentiWS–a Publicly Available German-language Resource for Sentiment Analysis. *In Proceedings of the*

7th International Language Resources and Evaluation, pages 1681–1711. Valletta, Malta.

Rice, J. A. 2006. Mathematical Statistics and Data Analysis. Duxbury, Massachusetts: Duxbury Advanced Series.

Riloff, E., and Lehnert, W. 1994. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, 12(3):296–333.

Riloff, E., Wiebe, J., and Wilson, T. 2003. Learning Subjective Nouns using Extraction Pattern Bootstrapping. *In Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003*, pages 25–32. Edmonton, Canada.

Rosso, P., Ferretti, E., Jiménez, D., and Vidal, V. 2004. Text Categorization and Information Retrieval Using Wordnet Senses. *In Proceedings of the 2nd Global Wordnet Conference*, pages 299–304. Brno, Czech Republic.

Saad, Y. 2003. Iterative Methods for Sparse Linear Systems (the second edition). Society for Industrial and Applied Mathematics.

Sarawagi, S. 2008. Information Extraction. *Foundations and Trends in Databases*, 1(3):261–377.

Sauper, C., Haghighi, A., and Barzilay, R. 2011. Content Models with Attitude. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 350–358. Portland, Oregon, USA.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

Seki, Y., Evans, D. K., Ku, L. W., Sun, L., Chen, H. H., and Kando, N. 2008.

Overview of multilingual opinion analysis task at NTCIR-7. *In Proceedings of the NTCIR07 workshop*, pages 104–111. Tokyo, Japan.

Seki, Y., Ku, L. W., Sun, L., Chen, H. H., and Kando, N. 2009. Overview of Multilingual Opinion Analysis Task at NTCIR-8: A Step Toward Cross Lingual Opinion Analysis. *In Proceedings of the NTCIR08 workshop*, pages 209–220. Tokyo, Japan.

Somasundaran, S., and Wiebe, J. 2009. Recognizing stances in online debates. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234. Suntec, Singapore.

Steinberger, J., Lenkova, P., Ebrahim, M., Ehrmann, M., Kabadjov, A. H. M., Steinberger, R., and Vazquez, S. 2011. Creating Sentiment Dictionaries via Triangulation. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 28–36. Portland, Oregon, USA.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. 1963. The General Inquirer: A Computer Approach to Content Analysis. *In Proceedings of the spring joint computer conference,* pages 241–256. MIT Press.

Szymański, J., and Duch, W. 2012. Annotating Words Using WordNet Semantic Glosses. *In Proceeding of the 19th International Conference on Neural Information Processing*, pages 180–187. Doha, Qatar.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. 2011. Lexicon-based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Tackstrom, O., and McDonald, R. 2011. Discovering Fine-grained Sentiment with Latent Variable Structured Prediction Models. *In Proceedings of the 33rd European Conference on Advances in Information Retrieval*, pages 368–374. Dublin, Ireland.

Takamura, H., Inui, T., and Okumura, M. 2005. Extracting semantic orientations of words using spin model. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 130–140. Ann Arbor, Michigan.

Téllez Valero, A., Montesy Gómez, M., and Villasenor-Pineda, L. 2005. A Machine Learning Approach to Information Extraction. *Computational Linguistics and Intelligent Text Processing*, 3406:539–547.

Titov, I., and McDonald, R. 2008. Modeling Online Reviews with Multi-grain Topic Models. *In Proceedings of the 17th International Conference on World Wide Web*, pages 111–120. Beijing, China.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 178–185. Atlanta, Georgia, USA.

Turney, P. D. 2002. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Philadelphia, Pennsylvania.

Turney, P., and Littman, M. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21(4):315–346.

Wan, X. 2009. Co-training for cross-lingual sentiment classification. *In*

Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language, pages 235–243. Suntec, Singapore.

Wang, D., and Liu, Y. 2011. A Pilot Study of Opinion Summarization in Conversations. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 331–339. Portland, Oregon, USA.

Wang, H., Huang, H., and Ding, C. 2011. Image Annotation Using Bi-Relational Graph of Images and Semantic Labels. *In Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 793–800. Colorado Springs, CO, USA.

Whitelaw, C., Garg, N., and Argamon, S. 2005. Using appraisal groups for sentiment analysis. *In Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 625–631. Bremen, Germany.

Wiebe, J., and Riloff, E. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *In Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, pages 486–497. Mexico City, Mexico: Springer-Verlag.

Wiebe, J., Wilson, T., and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural*

Language Processing, pages 347–354. Vancouver, British Columbia, Canada.

Wu, Y., Zhang, Q., Huang, X., and Wu, L. 2009. Phrase dependency parsing for opinion mining. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1541. Singapore.

Xu, H., Zhao, K., Qiu, L., and Hu, C. 2010. Expanding Chinese sentiment dictionaries from large scale unlabeled corpus. *In Proceedings of the International conference on computational linguistics and related areas*, pages 301–310. Beijing, China.

Yano, T., and Smith, N. A. 2010. What's Worthy of Comment? Content and Comment Volume in Political Blogs. *In Proceedings of the 4th International Conference on Weblogs and Social Media*. Washington, DC, USA.

Yu, H., and Hatzivassiloglou, V. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129 – 136. Sapporo, Japan.

Yu, J., Zha, Z. J., Wang, M., Wang, K., and Chua, T. S. 2011. Domain-assisted Product Aspect Hierarchy Generation: Towards Hierarchical Organization of Unstructured Consumer Reviews. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 140–150. Edinburgh, United Kingdom.

Zha, Z. J., Mei, T., Wang, J., Wang, Z., and Hua, X. S. 2009. Graph-based Semi-supervised Learning with Multiple Labels. *Journal of Visual Communication and Image Representation*, 20(2):97–103.

Zhang, H. 2004. The Optimality of Naive Bayes. *In Proceeding of the 17th International FLAIRS Conference*. Miami Beach, Florida.

Zhang, W., Gao, D., and Li, W. 2010. PolyU at MOAT in NTCIR-8. *In Proceedings of NTCIR-8 Workshop Meeting*, pages 282–285. Tokyo, Japan.

Zhao, W. X., Jiang, J., Yan, H., and Li, X. 2010. Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65. Cambridge, Massachusetts, USA.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Scholkopf, B. 2003. Learning with Local and Global Consistency. *In Proceedings of the 17th Annual Conference on Neural Information Processing Systems*. British Columbia, Canada.

Zhu, X., and Ghahramani, Z. 2002. Learning from Labeled and Unlabeled Data with Label Propagation. CMU CALD tech report.

Zhu, X., Goldberg, A. B., Brachman, R., and Dietterich, T. 2009. Introduction to Semi-Supervised Learning Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.

Zhuang, L., Jing, F., and Zhu, X. Y. 2006. Movie review mining and summarization. *In Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 43–50. Arlington, Virginia, USA.