Emotion Detection in Code-switching Texts via Bilingual and Sentimental Information

Zhongqing Wang[†‡], **Sophia Yat Mei Lee**[‡], **Shoushan Li**[†*], **and Guodong Zhou**[†]
[†]Natural Language Processing Lab, Soochow University, China
[‡]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
{wangzq.antony, sophiaym}@gmail.com,
{lishoushan, gdzhou}@suda.edu.cn

Abstract

Code-switching is commonly used in the free-form text environment, such as social media, and it is especially favored in emotion expressions. Emotions in code-switching texts differ from monolingual texts in that they can be expressed in either monolingual or bilingual forms. In this paper, we first utilize two kinds of knowledge, i.e. *bilingual* and *sentimental* information to bridge the gap between different languages. Moreover, we use a term-document bipartite graph to incorporate both bilingual and sentimental information, and propose a label propagation based approach to learn and predict in the bipartite graph. Empirical studies demonstrate the effectiveness of our proposed approach in detecting emotion in code-switching texts.

1 Introduction

With the rapid development of Web 2.0, emotion analysis in social media has become of great value to market predictions and analysis (Liu et al., 2013; Lee et al., 2014). Previous researches on emotion analysis have mainly focused on emotion expressions in monolingual texts (Chen et al., 2010; Lee et al., 2013a). However, in informal settings such as micro-blogs, emotions are often expressed by a mixture of different natural languages. Such a mixture of language is called code-switching. Specifically, code-switching text is defined as text that contains more than one language (code). It is a common phenomenon in multilingual communities (Auer, 1999; Adel et al., 2013). For instance, [E1-E3] are three examples of code-switching emotional posts containing both Chinese and English words. [E1] expresses the *happiness* emotion through English, and the *anger* emotion in [E2] is expressed through both Chinese and English, while the *fear* emotion in [E3] is expressed through a mixed English-Chinese phrase (hold不住).

> [E1] 我们已经自**high**起来了
> (*We are already getting **hyper** ourselves.*)
>
> [E2] 最厌恶的一句话就是“爱情没有先来后到，不被爱的才是第三者”。**shit!**
> (*A quote, to my great disgust, is "There's no staking claims in a relationship based on who got there first - the one who isn't loved is the true third party." **Shit!***)
>
> [E3] 这么个划重点法。。。窝们**hold不住**啊！！！
> (*The so-called "highlighting"...we **can't hold it anymore***.)

It is more difficult to detect emotions in code-switching texts than in monolingual ones since emotions in code-switching posts can be expressed through one or two languages. Hence, traditional automatic emotion detection methods which simply consider monolingual texts (Liu et al., 2013; Lee et al., 2013a) would not be readily applicable.

The key issue of emotion detection in code-switching texts is to deal with the emotions expressed through different languages. Thus bridging the gap between different languages becomes essential for emotion detection in code-switching texts. A straightforward approach to handle this issue is to translate texts from one language into another. Since Chinese is the dominant language in our data set, a word-by-word statistical machine translation strategy (Zhao et al., 2009) is adopted to translate English words into Chinese. Additionally, as text from micro-blogs is informal,

*Corresponding author

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 763–768,
Beijing, China, July 26-31, 2015.

synonym dictionary and PMI similar based word correlation (Turney, 2002) are used to enhance the language model for machine translation.

In spite of the English-to-Chinese translation, many English and Chinese words are still unconnected. Hence, we use sentiment analysis strategy (Turney, 2002; Li et al., 2013) to extract the polarity of both Chinese and English texts, and then connect words of similar polarity.

Moreover, for propagating label information between the bilingual texts from training data to test data, we use a term-document bipartite graph to incorporate both bilingual and sentimental information and propose a label propagation (Zhu and Ghahramani, 2002) based approach to learn and predict in the graph. Specially, the label information between Chinese and English texts would be propagated through the bipartite graph by word-document relations, bilingual information, and sentiment information. Evaluation of the data set indicates the importance of the task and the effectiveness of our proposed approach.

2 Related Work

Emotion analysis has been a hot research topic in NLP in the last decade. One main group of related studies on this task is about emotion resource construction (Xu et al., 2010; Volkova et al., 2012; Lee et al., 2014). Moreover, emotion classification is one of the most important tasks in emotion analysis, while emotion classification aims to classify text into multiple emotion categories (Chen et al., 2010; Liu et al., 2013). Despite a growing body of research on emotion analysis, little has been done on the analysis of emotion in code-switching due to the complexities of processing two languages at the same time.

Besides, although several research studies have focused on analyzing bilingual (Wan, 2009; Lu et al., 2011; Tang et al., 2014) and code-switching texts (Li and Fung, 2012; Ling et al., 2013; Lignos and Marcus, 2013), none of them has studied the multilingual code-switching issues in emotion detection. This research area is especially crucial when public emotions are mostly expressed in the free-form text on the Internet.

3 Data Collection

We collect our data set from *Weibo.com*, one of the most popular SNS websites in China. We use encoding code for each character in the post to identify the code-switching posts. After removing posts containing noise and advertisements, we extract 4,195 code-switching posts from the dataset for emotion annotation. Five basic emotions are annotated, namely *happiness*, *sadness*, *fear*, *anger* and *surprise* (Lee et al., 2013b). After the annotation process, results show 2,312 posts which include emotions. Moreover, 81.4% of emotional posts are expressed through Chinese. Although there are a few words of English in each post (an average of 3 words per post), 43.5% of emotion posts are caused by English. This statistic indicates that English is of vital importance to emotion expression even in code-switching contexts dominated by Chinese.

The corpora is annotated by two annotators and the inter-annotator agreement calculation shows that the agreement of our annotation is 0.692 in Cohen's Kappa coefficient, which indicates that the quality of the annotation is guaranteed.

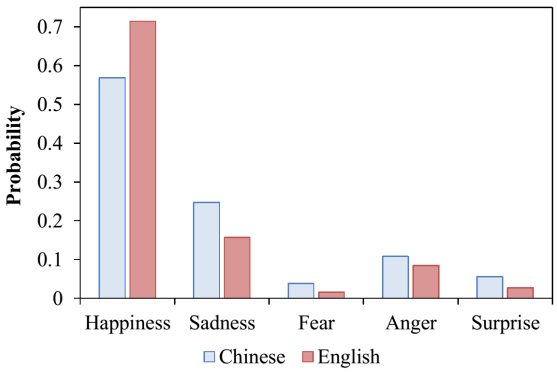

Figure 1: Distribution of Emotions and Languages

The joint distribution between emotions and caused languages is illustrated in Figure 1. The Y-axis of the figure presents the conditional probability of a post expressing the emotion e_i given that l_j is the caused language, $p(e_i|l_j)$.

It is suggested in Figure 2 that: 1) *happiness* occurs more frequently than other emotions; 2) people would like to use English text to express the *happiness* emotion much more than the *sadness* emotion; 3) the distribution of emotions expressed through Chinese and English text are similar.

4 Emotion Detection via Bilingual and Sentiment Information

In this paper, our goal is to predict the emotion label for each unlabeled post. Simply, we only choose those posts with single emotion on our re-

search. We systematically explore both the bilingual and sentimental information to detect emotions in code-switching posts. Moreover, we use a term-document bipartite graph to incorporate these two kinds of information, and propose a Label Propagation (LP) based approach to learn and predict emotion in code-switching texts. In the following subsections, we will discuss these issues one by one.

4.1 Bilingual Information

For using bilingual information, a word-by-word statistical machine translation strategy is adopted to translate words from English into Chinese. For better clarity, a word-based decoding, which adopts a log-linear framework as in (Och and Ney, 2002) with translation model and language model being the only features, is used:

$$P(c|e) = \frac{\exp\left[\sum_{i=1}^{2} \lambda_i h_i(c,e)\right]}{\sum_c \exp\left[\sum_{i=1}^{2} \lambda_i h_i(c,e)\right]} \quad (1)$$

where

$$h_1(c,e) = \log(p_\gamma(c|e)) \quad (2)$$

is the translation model, which is converted from the bilingual lexicon[1], and

$$h_2(c,e) = \log(p_{\theta_{LM}}(c) p_{\theta_{SYN}}(c) p_{\theta_{PMI}}(c)) \quad (3)$$

is the language model, and $p_{\theta_{LM}}(c)$ is the bigram language model which is trained from a large scale *Weibo* data set[2]. As text in micro-blogs is informal, synonym dictionary[3] and PMI based word correlation are used to enhance the language model for machine translation. $p_{\theta_{SYN}}(c)$ denotes the synonym similarity between translated words and the contexts. This is necessary since the sense of translated words and the contexts are expected to be similar; and $p_{\theta_{PMI}}(c)$ presents the PMI similarity between translated words and the contexts, while the PMI score is calculated by the individual and co-occurred hit count between translated words and contexts from the search engine[4] (Turney, 2002). This is to ensure that the translated words are highly associated with the contexts.

The candidate target sentences made up of a sequence of the optional target words are ranked by the language model. The output will be generated only if it reaches the maximum probability as follows (Brown et al., 1990; Zhao et al., 2009):

$$c = argmax \prod p(w_c) \quad (4)$$

4.2 Sentimental Information

Sentimental information is very useful in emotion detection (Gao et al., 2013). In this paper, we extract polarity from both Chinese and English texts to ensure text of similar polarity will be connected.

In this paper, both Chinese[5] and English[6] sentimental lexicons are employed to identify candidate opinion expressions by searching the occurrences of negative and positive expressions in text, and predict the polarity of both Chinese and English texts through the word-counting approach (Turney, 2002).

4.3 LP-based Emotion Detection

For the knowledge of bilingual and sentimental information to be well incorporated, we use a term-document bipartite graph to incorporate the information, and propose a label propagation based approach to learn and predict emotion in code-switching texts.

The input of the LP algorithm is a graph describing the relationship between each sample pair in the labeled and test data (Sindhwani and Melville, 2008; Li et al., 2013). In a bipartite graph, the nodes consist of two parts: documents and all terms extracted from the documents. An undirected edge (d_i, w_k) exists if and only if the document d_i contains the term w_k.

Note that, there are four kinds of terms on the graph, i.e., Chinese words, English words, translated Chinese words (bilingual information), and sentimental features. Although Chinese words and English words cannot be connected directly, the label information between Chinese and English words would be propagated through the bipartite graph by word-document relations, bilingual information, and sentiment information. The example of the bipartite graph is illustrated on the Figure 2.

[1] *MDBG CC-CEDICT* is adopted as the bilingual lexicon: http://www.mdbg.net/chindict/chindict.php?page=cedict

[2] The large-scale *Weibo* data set contains 2,716,197 posts in total.

[3] *TongYiCiLin* is adopted as the Chinese synonym dictionary: http://www.ltp-cloud.com/

[4] We use *BING.com* as the search engine for PMI: http://www.bing.com/

[5] *DUTIR Sentiment Lexicon* is adopted as the Chinese sentiment lexicon: http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx

[6] English sentiment lexicon is utilized from *MPQA Subjectivity Lexicon*: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

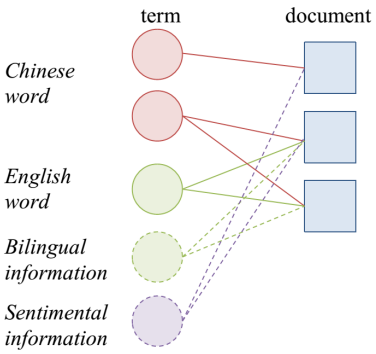

Figure 2: Example of the bipartite graph

When all terms are taken into consideration, we get the transition probability from d_i to d_j as in (5):

$$t_{ij} = \sum_k \frac{x_{ik}}{\sum_k x_{ik}} \cdot \frac{x_{jk}}{\sum_k x_{jk}} \quad (5)$$

where x_{ik} is the frequency of term w_k in document d_i.

After building the document-document transfer matrix through the bipartite graph, we use label propagation algorithm (Zhu and Ghahramani, 2002; Zhou and Kong, 2009) to learn and predict emotions in the graph, in which the probabilities of the labeled data are clamped in each loop using their initial ones and act as a force to propagate their labels to the test data.

5 Experiments

In this section, we first introduce the experimental settings, and then evaluate the performance of our proposed approach for detecting emotions in code-switching texts.

5.1 Experimental Settings

As described in Section 3, the data are collected from *Weibo.com*. We randomly select half of the annotated posts as the training data and another half as the test data. FNLP[7] is used for Chinese word segmentation.

5.2 Experimental Results

Our first group of experiments is to investigate whether our proposed label propagation model with both bilingual and sentimental information can improve emotion detection in code-switching texts. Figure 3 shows the experimental results of different models, where *ME* is the basic Maximum Entropy (ME) classification model[8] in which all Chinese and English words of each post function as a feature, *ME-CN* and *ME-EN* in which only the Chinese or English text of each post function as features, and *BLP-BS*, our proposed LP-based approach which incorporates both bilingual and sentimental information. We adopt F1-Measure (F1.) to measure the performance of each model in the respective emotions.

From Figure 3, we find that the results of *ME-CN* and *ME-EN* are instable. It indicates that only considering one kind of language text is not very effective for predicting emotions in code-switching texts. Moreover, as Chinese and English texts are taken into account collectively with both bilingual and sentimental information, our proposed *BLP-BS* model is significantly better than basic approaches on all the emotions.

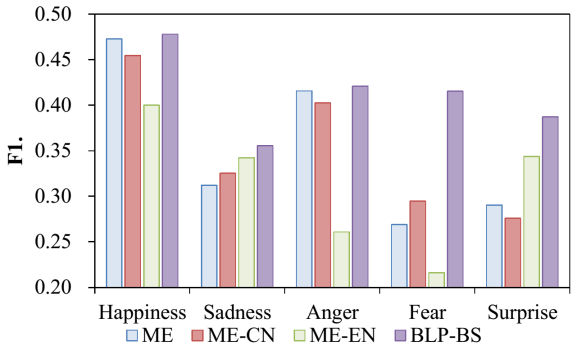

Figure 3: Results of emotion detection

We then analyze the influence of different factors in our proposed approach with average F1-Measure of the five emotions with the results illustrated in Table 1. In the table, *Basic SMT* refers to using basic word-by-word statistical machine translation to help the detection process; *Enhanced SMT* refers to using both synonyms and word correlation to enhance the machine translation process; *Sentiment* refers to using sentimental information to help the detection process; *ME-BS* refers to using the maximum entropy model with both bilingual and sentimental information, and *BLP* refers to the label propagation model in which all of the words in Chinese and English text function as a feature.

From Table 1, it is observed that: 1) sentimental information (*Sentiment*) are effective for predicting emotion in both *ME*-based and *BLP*-

[7] *FNLP (FudanNLP)*, https://github.com/xpqiu/fnlp/

[8] ME algorithm is implemented with the *MALLET Toolkit*, http://mallet.cs.umass.edu

Method	Average F1.
ME	0.354
+Basic SMT	0.354
+Enhanced SMT	0.382
+Sentiment	0.369
ME-BS	0.383
BLP	0.385
+Enhanced SMT	0.392
+Sentiment	0.406
BLP-BS	**0.412**

Table 1: Results of influence on different factors

based models; 2) *Enhanced SMT* outperforms *Basic SMT*, which proves the effectiveness of our enhanced approaches for statistical machine translation; and 3) our proposed approach (*BLP-BS*) outperforms the other approaches. This indicates the complementarity of bilingual and sentimental information on the bipartite graph based label propagation model.

6 Conclusion

In this study, we address a novel task, namely emotion detection in code-switching texts. First, we collect and extract the code-switching posts from *Weibo.com*, which are annotated with emotions. Then, we use both SMT-based bilingual information and sentimental information to bridge the gap between different languages in code-switching texts. Finally, we propose a bipartite graph based label propagation model to effectively incorporate both bilingual and sentimental information for detecting emotion in code-switching texts. Empirical studies demonstrate that our model significantly outperforms several strong baselines.

Our current work assumes the independence of emotions and caused languages. In future work, we would like to explore the relation among emotions and caused languages for detecting the emotion and caused languages collectively.

Acknowledgments

Prof. Deyi Xiong contributed valuable insights and wise counsel on machine translation. Our colleague, Helena Yan Ping Lau, did excellent work on corpus analysis for probing and understanding the problem, and helped us review the paper, offering valuable feedback and helpful leads.

We thank our anonymous reviewers for prudent advice.

The work is funded by an Early Career Scheme (ECS) sponsored by the Research Grants Council of Hong Kong (No. PolyU 5593/13H), and supported by the National Natural Science Foundation of China (No. 61273320, and No. 61375073) and the Key Project of the National Natural Science Foundation of China (No. 61331011).

References

Adel H., N. Vu, and T. Schultz. 2013. Combination of Recurrent Neural Networks and Factored Language Models for Code-Switching Language Modeling. In *Proceedings of ACL-13*.

Auer P. 1999. *Code-Switching in Conversation*. Routledge.

Brown P., J. Cocke, S. Pietra, V. Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79-85.

Chen Y., S. Lee, S. Li, and C. Huang. 2010. Emotion Cause Detection with Linguistic Constructions. In *Proceeding of COLING-10*.

Gao W., S. Li, S. Lee, G. Zhou, and C. Huang. 2013. Joint Learning on Sentiment and Emotion Classification. In *Proceedings of CIKM-13*.

Lee S., H. Zhang, and C. Huang. 2013a. An Event-Based Emotion Corpus. In *Proceedings of CLSW 2013*.

Lee S., Y. Chen, C. Huang, and S. Li. 2013b. Detecting Emotion Causes with a Linguistic Rule-Based Approach. *Computational Intelligence*, 29(3), 390-416.

Lee S., S. Li, and C. Huang. 2014. Annotating Events in an Emotion Corpus. In *Proceedings of LREC-14*.

Li Y., and P. Fung. 2012. Code-switch Language Model with Inversion Constraints for Mixed Language Speech Recognition. In *Proceedings of COLING-12*.

Li S., Y. Xue, Z. Wang, and G. Zhou. 2013. Active Learning for Cross-domain Sentiment Classification. In *Proceeding of IJCAI-2013*.

Ling W., G. Xiang, C. Dyer, A. Black, and I. Trancoso. 2013. Microblogs as Parallel Corpora. In *Proceedings of ACL-13*.

Lignos C., and M. Marcus. 2013. Toward Web-scale Analysis of Codeswitching. In *Proceedings of Annual Meeting of the Linguistic Society of America*.

Liu H., S. Li, G. Zhou, C. Huang, and P. Li. 2013. Joint Modeling of News Reader's and Comment Writer's Emotions. In *Proceedings of ACL-13*, shorter.

Lu B., C. Tan, C. Cardie and B. Tsou. 2011. Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. In *Proceedings of ACL-2011*.

Och F., and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of ACL-02*.

Quan C., and F. Ren. 2009. Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis. In *Proceedings of EMNLP-09*.

Sindhwani V. and P. Melville. 2008. Document-Word Co-Regularization for Semi-supervised Sentiment Analysis. In *Proceedings of ICDM-08*.

Tang X., X. Wan, and X. Zhang. 2014. Cross-Language Context-aware Citation Recommendation in Scientific Articles. In *Proceedings of SIGIR-14*.

Turney P. 2002. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of comments. In *Proceedings of ACL-02*.

Volkova S., W. Dolan, and T. Wilson. 2012. CLex: A Lexicon for Exploring Color, Concept and Emotion Associations in Language. In *Proceedings of EACL-12*.

Xu G., X. Meng, and H. Wang. 2010. Build Chinese Emotion Lexicons Using A Graph-based Algorithm and Multiple Resources. In *Proceeding of COLING-10*.

Wan X. 2009. Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of ACL/IJCNLP-09*.

Zhao H., Y. Song, C. Kit, and G. Zhou. 2009. Cross Language Dependency Parsing using a Bilingual Lexicon. In *Proceedings of ACL-09*.

Zhou G. and K. Fang. 2009. Global Learning of Noun Phrase Anaphoricity in Coreference Resolution via Label Propagation. In *Proceedings of EMNLP-2009*.

Zhu X. and Z. Ghahramani. 2002. Learning from Labeled and Unlabeled Data with Label Propagation. *CMU CALD Technical Report*. CMU-CALD-02-107.