Multi-view Learning for Emotion Detection in Code-switching Texts

Sophia Yat Mei Lee[†], and Zhongqing Wang[†,‡]

[†] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

[‡] Natural Language Processing Lab, Soochow University, China

{sophiaym, wangzq.antony}@gmail.com

***Abstract*—Previous researches have placed emphasis on analyzing emotions in monolingual text, neglecting the fact that emotions are often found in bilingual or code-switching posts in social media. Traditional methods for the identification or classification of emotion fail to accommodate the code-switching content. To address this challenge, in this paper, we propose a multi-view learning framework to learn and detect the emotions through both monolingual and bilingual views. In particular, the monolingual views are extracted from the monolingual text separately, and the bilingual view is constructed with both monolingual and translated text collectively. Empirical studies demonstrate the effectiveness of our proposed approach in detecting emotions in code-switching texts.**

Keywords—emotion analysis; code-switching; multi-view learning

I. INTRODUCTION

Due to the popularity of opinion-rich resources (e.g., online review sites, forums, and micro-blog websites), emotion analysis in text has become of great significance in obtaining useful information for studies in social media (Pang et al., 2002; Liu et al., 2013; Lee et al., 2014). Previous researches have focused on analyzing emotions in monolingual texts (Chen et al., 2010; Lee et al., 2013a). However, code-switching posts are common in social media, where emotions can be expressed in either monolingual or bilingual forms. Here, code-switching text is defined as text that contains more than one language ('code') (Adel et al., 2013; Auer, 1999). Below are three examples of code-switching posts on Weibo.com that contain both Chinese and English texts. While **Example 1** expresses the *happiness* emotion in English, and **Example 2** expresses the *sadness* emotion in both Chinese and English, the *sadness* emotion in **Example 3** is expressed in a mixed Chinese-English phrase (hold 不住, '*cannot take it*').

Example 1: 婚礼上新娘大秀歌技， **high** 翻全场！
(*Bride sings on the wedding, all the people are getting* ***hyper****!*)

Example 2: shit，失眠了。明明很困却睡不着，难受。
(***Shit****, suffering from insomnia. Feeling so drowsy but can't sleep, I'm distressed.*)

Example 3: 聊天聊到早上六点，我 **hold** 不住了，好困好困。
(*Have been chatting till 6 a.m., I* ***can't take it anymore****, so very sleepy.*)

From the above examples, we find that it is much more difficult to detect emotions in code-switching text than in monolingual text, since the emotions in code-switching posts could be expressed in either one (**Example 1**) or two languages (**Example 2** and **Example 3**). The traditional methods, which only consider monolingual texts, would be inadequate for code-switching texts. Hence, it is necessary to learn the detection model from both monolingual and bilingual text collectively. To address this challenge, a multi-view based semi-supervised learning framework is utilized to learn and detect the emotion by both monolingual and bilingual views. In particular, the Chinese and English texts are employed to build the two monolingual views (Chinese view and English view). Moreover, as the monolingual views are employed to learn from the monolingual text separately, we use a statistical machine translation strategy (Zhao et al., 2009) to translate English text into Chinese to build the bilingual view. In addition, both sentiment and synonym information are used to enhanced the bilingual view. Finally, a co-training approach is utilized to incorporate both monolingual and bilingual views. Experimental results show that our approach is effective, and apparently superior to using multi-view learning for detecting emotions in code-switching texts.

The remainder of this paper is organized as follows: Section 2 provides an overview of the related work on emotion analysis; Section 3 provides a description of the data and statistics; Section 4 presents the approach for detecting emotions in code-switching texts by multi-view learning; Section 5 discusses the empirical study; and Section 6 draws the conclusion and outlines future work.

II. RELATED WORK

In this section, we discuss related work on emotion analysis and code-switching texts.

Emotion analysis has been well studied in the community of natural language processing, with focus on lexicon building and emotion classification. On lexicon building, Rao et al. (2012) automatically built a word-emotion mapping dictionary for social emotion detection. Yang et al., (2014) proposed an emotion-aware topic model to build a domain specific lexicon. On emotion classification, Liu et al., (2013) proposed a co-training framework to infer the news reader's and comment writer's emotions collectively. Wen and Wan (2014) used class sequential rules for emotion classification of micro-blog texts by regarding each post as a data sequence.

Research on code-switching texts can be traced back to the 1970s. Since then, several theories, such as diglossia (Blom and Gumperz, 1972), the communication accommodation theory (Giles and Clair, 1979), the markedness

model (Myers-Scotton, 1993), and the conversational analysis model (Auer, 1984), have been proposed to account for the motivation behind code-switching. Meanwhile, code-switching documents have also received considerable attention in the NLP community, with a focus on identification and analysis, including mining translations in code-switching documents (Ling et al., 2013), predicting code-switching points (Solorio and Liu, 2008), identifying code-switching tokens (Lignos and Marcus, 2013), adding code-switching support to language models (Li and Fung, 2012), and learning poly-lingual topic models from code-switching text (Peng et al., 2014).

With the increasing popularity of multilingual natural language processing (due to its broad real-world applications), research on code-switching texts has been drawing more and more attention in various NLP and related tasks, such as parsing (Burkett et al., 2010), information retrieval (Gao et al., 2009), and sentiment analysis (Lu et al., 2011).

However, none have studied the multilingual code-switching issues in emotion analysis, which have become more and more crucial as our society has become more global and the public tends to express emotions on the Internet (with the arrival/ development of Web 2.0.)

III. Data Collection

We retrieve our data set from Weibo.com, one of the famous SNS websites in China. The encoding code for each character in the post is used to identify those code-switching posts. After removing those posts containing noise and advertisements, 4,195 code-switching posts are extracted for emotion annotation. Identical to Lee et al. (2013b), five basic emotions are annotated, namely *happiness, sadness, fear, anger and surprise*.

The data is annotated by two annotators with the inter-annotator agreement reaching 0.692 in Cohen's Kappa coefficient. This indicates that annotation quality is guaranteed. Out of 4,195 annotated posts, 2,312 posts are of emotions. Moreover, 81.4% of emotional posts are expressed in Chinese text. Although there are relatively fewer words in English contained in each post, 43.5% of emotional posts are expressed in English. This statistic shows that English is of vital importance to emotion expression even in code-switching contexts dominated by Chinese. There are overlaps between Chinese and English emotional posts, since some of the emotions are conducted in both Chinese and English in the same post.

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, $p(e_i \mid l_j)$. The figure shows that: 1) *happiness* occurs more frequently than the other emotions; 2) people tend to use English in expressing the *happiness* emotion than the *sadness* emotion; 3) the distribution of emotions expressed in Chinese and English text are similar.

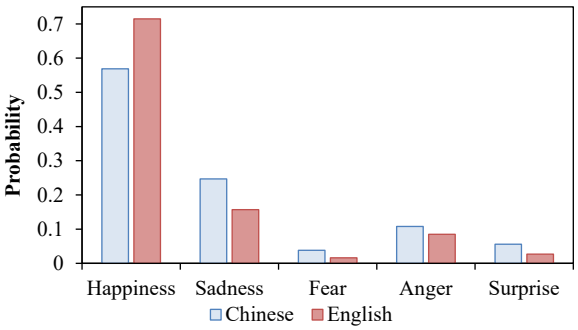

Figure 1. Joint Distribution of Emotions and Caused Languages

IV. Multi-View Learning For Emotion Detection

Our approach first generates monolingual and bilingual views from both Chinese and English text collectively from all code-switching posts. We then use co-training algorithm with the obtained views to perform semi-supervised learning for detecting emotions. Figure 2 demonstrated the framework of our approach.

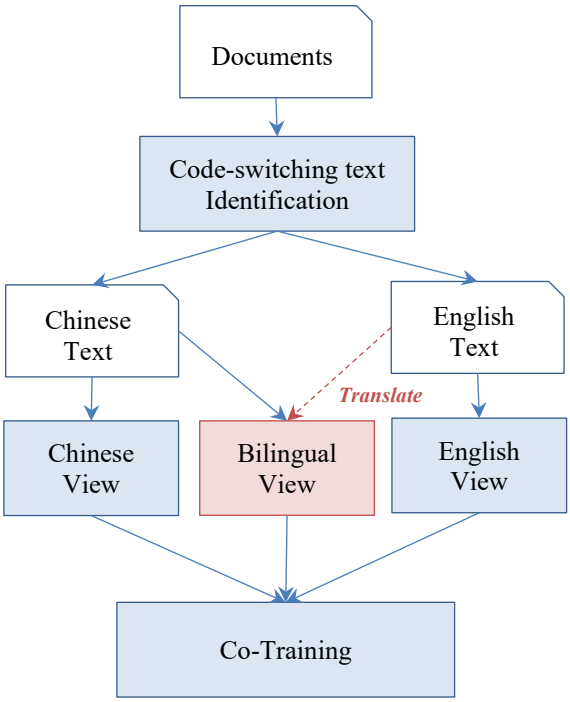

Figure 2. The Overview of Our Approach

From Figure 2, we find that there are three views in our framework, of which the *Chinese View* is extracted from the Chinese text of each post, the *English View* is extracted from the English text of each post, and the *Bilingual View* is translated from English text and combined with Chinese text. In the following parts, we will discuss the bilingual view in detail, and describe the co-training algorithm of our study.

A. Bilingual View

As the monolingual views use the Chinese or English text separately, a bilingual view is employed to combine Chinese and English text in a single view. 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 to build

the Bilingual View. In addition, both sentiment information and synonym information are used to enhance the translated view.

Firstly, a word-by-word statistical machine translation strategy is adopted to translate words from English into Chinese. To be more specific, 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:

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

where

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

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

$$h_2(c,e) = \log(p_\theta(c)) \quad (3)$$

is the language model, and $p_\theta(c)$ is the bigram language model which is trained from a large scale Weibo data set[2].

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):

$$c = \arg\max \prod p(w_c) \quad (4)$$

Moreover, two kinds of information are used to enhance the bilingual view:

- **Sentimental information** is very useful in emotion detection (Gao et al., 2013). In this paper, we extract polarity from both Chinese and English text to ensure text of similar polarity will be connected. Both Chinese[3] and English[4] 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 text through the word-counting approach (Turney, 2002).
- **Synonym** is also utilized to combine similar Chinese with English words in the bilingual view. In particular, a Chinese synonym dictionary[5] is used to find synonym relation between translated words and original Chinese words. This is necessary since the sense of translated words and the contexts are expected to be similar.

B. Co-Training

First of all, three different views are generated through both Chinese and English text, as shown in Figure 3. Given the three views, our approach for semi-supervised emotion detection in code-switching text is to use them to perform co-training algorithm (Blum and Mitchell, 1998; Wan, 2009; Liu et al., 2013), as shown in Figure 3. In our implementation, we set *n*=2.

In the training phase, three separate classifiers: C_{CN}, C_{EN}, and C_{BI} are obtained. Accordingly, in the classification phase, we then obtain three prediction values for a test post. To make the final decision, the average of the three prediction values is used.

ALGORITHM 1: CO-TRAINING

Given

- F_{CN}, F_{EN}, and F_{BI} are redundantly sufficient sets of features generated, where F_{CN} represent the features of Chinese View, F_{EN} represent the features of English View, and F_{BI} represents the Bilingual View.
- L is a set of labeled training posts.
- U is a set of unlabeled posts.

Loop for *N* iteration:

1) Learn the classifier C_{CN} from L based on F_{CN};
2) Use C_{CN} to label reviews from U based on F_{CN};
3) Choose n most confidently predicted reviews E_{CN} from U;
4) Similar to step 1), 2), and 3), learn classifier C_{EN}, C_{BI} respectively and choose posts E_{EN}, E_{BI} from U;
5) Remove posts E_{CN}, E_{EN}, E_{BI} from U;
6) Add posts E_{CN}, E_{EN}, E_{BI} with the corresponding labels to L;

Figure 3. The Co-training Algorithm

V. EXPERIMENTS

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

A. Experimental Settings

As described in Section 3, the data are collected from Weibo.com. We randomly select 20% of the annotated posts as the labeled data, 60% as the unlabeled data, and the remaining as the testing data. *FNLP*[6] is used for Chinese word segmentation, and average F1-Measure (F1.) of all emotions is adopted to measure the performance.

B. Experimental Results

For thorough comparison, the following methods are implemented including:

- **Baseline**: only the labeled training data is used for training (with no unlabeled data), i.e., supervised

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

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

[5] *TongYiCiLin (同义词林)* is adopted as the Chinese synonym dictionary: http://www.ltp-cloud.com/

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

learning. Maximum Entropy (ME) classification model[7] is used as the basic classification model.

- **ME-CN**: only the Chinese text of each post is used as a feature to train a ME classification model.
- **ME-EN**: only the English text of each post is used as a feature to train a ME classification model.
- **Self-Training**: all the text of each post is used as a single view to training a self-training model.
- **Co-Training**: a co-training algorithm with both Chinese and English views is used.
- **Multi-View Learning:** use co-training algorithm with both Monolingual Views (Chinese view and English view) and bilingual view.

TABLE I. COMPARE WITH BASELINES

	Average F1.
Baseline	0.465
ME-CN	0.425
ME-EN	0.325
Self-Training	0.463
Co-Training	0.472
Multi-View Learning	**0.486**

From Table 1, we find that: 1) the performance of the basic approach (Baseline) which uses mixed text directly is inferior; 2) as Chinese is the dominant language, and the English text is loosely distributed, using Chinese text (ME-CN) outperforms using English text (ME-EN). Besides, as the English texts in the posts are always composed of single words, the average F1-score of ME-EN is much lower than the other two supervised approaches; 3) due to the incorporation of both Chinese and English text in a co-training framework, the Co-Training approach outperforms the basic supervised learning approaches; 4) by considering both monolingual and bilingual views, our proposed Multi-View Learning approach achieves a better performance than the other approaches. It also indicates that both monolingual and bilingual information in code-switching posts are effective for detecting emotions.

VI. 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 construct both monolingual and bilingual views to represent the texts of each post. Finally, we propose a multi-view learning framework to effectively incorporate both monolingual and bilingual information for detecting emotion in code-switching texts. Empirical studies demonstrate that our model significantly outperforms several strong baselines.

VII. ACKNOWLEDGMENTS

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).

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

REFERENCES

[1] 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.

[2] Auer P. 1999. Code-Switching in Conversation. Routledge.

[3] Blum A., and T. Mitchell. 1998. Combining Labeled and Unlabeled Data with Cotraining. In Proceedings of COLT-98.

[4] 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.

[5] Burkett, D., and D. Klein. 2008. Two Languages are Better than One (for Syntactic Parsing). In Pro-ceedings of EMNLP-08.

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

[7] Gao W., J. Blitzer, M. Zhou, and K. Wong. 2009. Ex-ploiting Bilingual Information to Improve Web Search. In Proceedings of ACL/IJCNLP-09.

[8] Gao W., S. Li, S. Lee, G. Zhou, and C. Huang. 2013. Joint Learning on Sentiment and Emotion Classification. In Proceedings of CIKM 2013.

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

[10] 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.

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

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

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

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

[15] 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.

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

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

[18] Rao Y., X. Quan, W. Liu, Q. Li, and M. Chen. 2012. Building Word-emotion Mapping Dictionary for Online News. In Proceedings of SDAD 2012 The 1st International Workshop on Sentiment Discov-ery from Affective Data.

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

[20] 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.

[21] 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.

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

[23] Wen S. and X. Wan. 2014. Emotion Classification in Microblog Texts Using Class Sequential Rules. In Proceedings of AAAI-14.

[24] Yang M., B. Peng, Z. Chen, D. Zhu, and K. Chow. 2014. A Topic Model for Building Fine-grained Domain-specific Emotion Lexicon. In Proceedings of ACL-14.

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