A Bilingual Attention Network for Code-switched Emotion Prediction

Zhongqing Wang[†], Yue Zhang[†], Sophia Yat Mei Lee[‡], Shoushan Li[§] and Guodong Zhou[§]
[†] Singapore University of Technology and Design, Singapore
[‡]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
[§] Natural Language Processing Lab, Soochow University, China
wangzq.antony@gmail.com, yue_zhang@sutd.edu.sg
ym.lee@polyu.edu.hk, {lishoushan, gdzhou}@suda.edu.cn

Abstract

Emotions in code-switching text can be expressed in either monolingual or bilingual forms. However, relatively little research has placed emphasis on code-switching text. The challenges of this task include the exploration both monolingual and bilingual information of each post and capturing the informative words from the code-switching context. To address these challenges, we propose a Bilingual Attention Network (BAN) model to aggregate the monolingual and bilingual informative words to form vectors from the document representation, and integrate the attention vectors to predict the emotion. The experiments show the effectiveness of the proposed model. Visualization of the attention layers illustrates that the model selects informative words qualitatively.

1 Introduction

Microblogs such as Twitter and Facebook have gained tremendous popularity in the past decade, they often contain extremely current, even breaking, information about world events. However, the writing style of microblogs tends to be quite colloquial and nonstandard, unlike the style found in more traditional, edited genres (Li et al., 2015; Vo and Zhang, 2015). In addition, authors from multi-lingual communities tend to write code-switching posts frequently (Ling et al., 2013; Wang et al., 2015). These pose challenges for automatic emotion prediction tasks.

There has been some previous research focusing on both emotion analysis (Pang et al., 2002; Lee et al., 2014) and code-switching text analysis (Solorio and Liu, 2008; Ling et al., 2013; Jamatia et al., 2015). However, little research has focused on predicting emotion in code-switching text. Different from monolingual emotion prediction, the emotion in code-switching posts can be expressed in either monolingual or bilingual forms. In this study, we focus on Chinese and English mixed code-switching text from Chinese social media. Although Chinese is the major language, it has been shown that English words are critical for emotion expression (20.1% posts express emotion through English text). E1 - E4 show four examples of code-switching posts that contain both Chinese and English words.

E 1. 玩了一下午轮滑**so happy**！
(I went rollerblading the whole afternoon, **so happy**!)

E 2. 开学以来，浮躁的情绪。不安稳的心态。确实该自己检讨一下了。。。**sigh**˜
(I have been grumpy and emotional since the first day of school, unstable mindset too. It's really time to self-evaluate ... **sigh**.)

E 3. 上了一天的课，嗓子**hold**不住了啊。
(I have been teaching the whole day, my throat **can't take it anymore**.)

E 4. 早起直接飙酒，喝多上车回校，回校一睁眼过站，多么**happy**的一天。
(I drank too much in the morning. I got drunk and went back to school by bus, and I missed my stop. Such a **happy** day.)

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1624–1634, Osaka, Japan, December 11-17 2016.

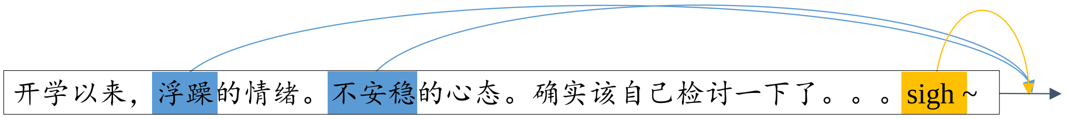
开学以来，浮躁的情绪。不安稳的心态。确实该自己检讨一下了。。。sigh ~

Figure 1: Example of attention mechanism result on E2.

These examples show that it is much more difficult to detect emotions in code-switching text than in the monolingual one, since emotions in code-switching posts can be expressed in either one or two languages. Take E2 for example, the sadness emotion is expressed by both Chinese and English text, and the mention of explicit emotion in English is triggered by the Chinese text. The sadness emotion of E3 is expressed by the mixed Chinese and English phrase "hold 不住". In addition, not all the words in the post are useful for predicting emotion, and the importance of words is highly context-dependent. For example in E2, only the words "浮躁" (*grumpy*) and "sigh" can be used to indicate the sadness emotion, yet the word "happy" in E4 indicates the opposite emotion without the context. Hence, the question how to explore both monolingual and bilingual information of each post post, and how to the informative words and phrases from the code-switching context are what constitute the challenging part.

To address the above challenges, we propose a Bilingual Attention Network (BAN) model to capture informative monolingual and bilingual emotion representations in the code-switching text. The attention mechanism (Bahdanau et al., 2014; Rocktäschel et al., 2015) used to aggregate the representation of informative words into a vector for emotion prediction, provides insight into which words contribute to the classification decision. In addition, a bilingual attention network is used to capture informative words from both monolingual and bilingual context. In particular, we first construct a document representation through a neural network model. Secondly, we project the document representation into three attention vectors by aggregating the representation of the informative words from both monolingual and bilingual context. Finally, a full-connected layer is used to integrate the three attention vectors and predict the emotion.

Our primary contribution is multi-lingual context sensitivity. The model considers both monolingual context and bilingual context, which allow the model to pay relevant attention to informative monolingual and bilingual words, respectively, when constructing relevant document representation. For example, consider the example in Figure 1. We can find that the informative Chinese and English words, such as "浮躁" (*grumpy*) and "sigh", have much more influence to the final decision. The key difference to previous work is that our system uses context to discover when a sequence of tokens is relevant rather than simply filtering for sequences of tokens, taken out of context. Evaluation shows the effectiveness of our proposed BAN model with both monolingual and bilingual information.

2 Related Works

2.1 Emotion Analysis

Over the last decade, there has been much work exploring various aspects of emotion analysis (Wiebe et al., 2005). While most focused on analyzing emotions in monolingual text. Some of these studies emotion lexicon building, for example, Rao et al. (2012) automatically built a word-emotion mapping dictionary for social emotion detection, Yang et al. (2014) proposed a novel emotion-aware topic model to build a domain-specific lexicon. For emotion classification, Liu et al. (2013) used a co-training framework to infer the news from readers' comments and writers' emotions collectively. Wen and Wan (2014) used class sequential rules for emotion classification of microblog texts by regarding each post as a data sequence. Li et al. (2015) proposed a factor graph based framework to incorporate both label and context dependency for emotion classification.

Deep neural networks have been proved effectiveness for many NLP tasks, including sentiment and emotion analysis (Vo and Zhang, 2015; Zhang et al., 2015). dos Santos and Gatti (2014) proposed a character-based deep convolutional neural network to predict sentiment of short text. Tang et al. (2015) proposed a neural network model to learn vector-based document representation. Zhang et al. (2015)

employed a neural network based CRFs for extracting opinion targets on open domains. Most of the previous studies focused on monolingual text, while our proposed bilingual attention network model focuses on exploring the monolingual and bilingual information collectively, and we also propose a new architecture to capture informative words from monolingual and bilingual contexts with attention mechanisms.

2.2 Research on Code-switching and Bilingual Text

Code-switched documents have received considerable attention in the NLP community (Adel et al., 2013; Garrette et al., 2015). Several studies have focused on code-switching identification and analysis, including mining translations in code-switched documents (Ling et al., 2013), predicting code-switched points (Solorio and Liu, 2008), identifying code-switched tokens (Lignos and Marcus, 2013), adding code-switched support to language models (Li and Fung, 2012), and POS tagging for code-switching text (Jamatia et al., 2015). There is relatively little work focus on predicting emotion in code-switching text. Wang et al. (2015) proposed a machine translation based approach to predict emotion in code-switching text with various external resources. Our approach departs from the previous work that we model the task by considering monolingual and bilingual information in both lexical and document level with neural network model and attention mechanism, while previous research only focused on lexical-level bilingual information. In addition, we do not use any external resource, such as bilingual and sentiment dictionary, to train our model.

More remotely connected, multilingual natural language processing has attracted increasing attention in the computational linguistic community due to its broad real-world applications. Relevant studies have been reported in various natural language processing tasks, such as parsing (Burkett and Klein, 2008), information retrieval (Gao et al., 2009), text classification (Amini et al., 2010), and so on. There are a number of studies on predicting sentiment polarity through multilingual text. Wan (2009) incorporated unlabeled data in the target language into classifier with co-training to improve classification performance. Wei and Pal (2010) regarded cross-lingual sentiment classification as a domain adaptation task and applied structural correspondence learning (SCL) to tackle this problem. Their approach achieves a better performance than the co-training algorithm. More recently, Meng et al. (2012) employed the parallel corpus for cross-lingual sentiment classification. They explored the case when no labeled data is available in the parallel corpus. However, such multi-lingual models do not explicitly consider code-switching, since their data sets are always parallel corpus. As the two languages are mixed in the code-switching text without parallel, code-switching corpus is more difficult to process.

3 Bilingual Attention Network

Given a post X with T words ($X =< w_1, w_2, ..., w_T >$), where each word w_t is represented with a K-dimensional embedding (Mikolov et al., 2013), our goal is to predict emotions for each post. Formally, for the post X with the emotion e, we need an objective integer variable y ($y \in \{0, 1\}$) to define if the emotion e is expressed in the post or not. Note that, we have five emotions in our emotion scheme[1], and we build the binary classifier to predict each emotion of the post individually.

There are three phases for predicting code-switched emotion using our bilingual attention network. Firstly, we use a long short-term memory network to build a document representation for each post. Secondly, we project the document representation into three vectors by aggregating the representation of the informative words from both monolingual and bilingual context. Note that, the attention vectors from monolingual context only consider the corresponding monolingual words from the document representation, and the vectors from bilingual context consider all the words from the document representation. Thirdly, a full-connected layer is used to integrate the three attention vectors, and predict the emotion with softmax function. The overall architecture of the Bilingual Attention Network (BAN) is shown in Figure 2.

[1] The emotions includes, *happiness*, *sadness*, *anger*, *fear*, and *surprise*. Please refer to Section 4.1 for more details.

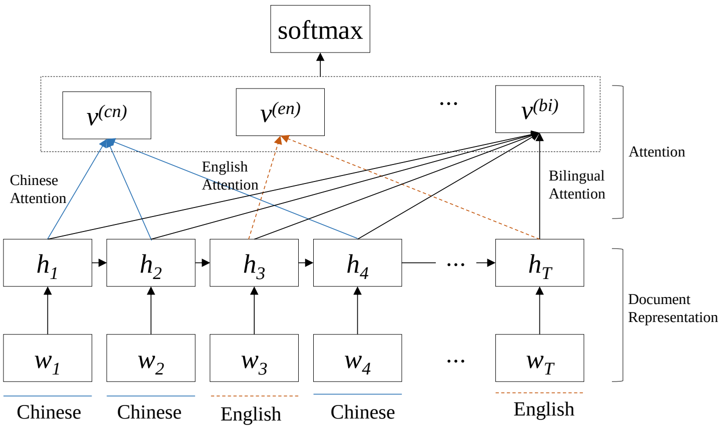

Figure 2: Overview of the bilingual attention network.

3.1 Document Representation

A Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) is used to obtain the document representation of each post. LSTM models a recurrent state transform sequence from a input sequence $\{x_1, x_2, ..., x_t\}$ of the post to a hidden state sequence $\{h_1, h_2, ..., h_t\}$. A LSTM represents each time step with an input, a memory and a output gate, denoted as i_t, f_t and o_t respectively. There are numerous variations to the LSTM model, and we choose one for which the hidden state h_t for each time-step t is given by:

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{i-1} + b^{(i)}) \tag{1}$$

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) \tag{2}$$

$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}) \tag{3}$$

$$u_t = \tanh(W^{(u)} + U^{(u)}h_{t-1} + b^{(u)}) \tag{4}$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1} \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where σ denotes the sigmoid function. After the LSTM process, we obtain an **annotation** h_t for a given word w_t.

3.2 Attention Mechanism

Not all words contribute equally to the representation of the meaning. Hence, we introduce an attention mechanism (Bahdanau et al., 2014; Yang et al., 2016) to extract the words that are important to the meaning of the post, and aggregate the representation of those informative words to form a vector. Since emotion can be expressed in either one or two languages in code-switching text, we build the vectors from monolingual and bilingual contexts respectively. For the monolingual case, we build two vectors $v^{(cn)}$ and $v^{(en)}$ to capture informative information from the Chinese and English contexts separately. For the bilingual case, we construct a vector $v^{(bi)}$ to capture the salient words from the mixed text.

Bilingual Attention. We use an attention function to aggregate the representation of the salient words to form the bilingual attention vector $v^{(bi)}$. Specifically,

$$u_t^{(bi)} = \tanh(W^{(bi)}h_t^{(bi)} + b^{(bi)}) \tag{7}$$

$$\alpha_t^{(bi)} = \frac{\exp(u_t^{(bi)\mathrm{T}}u^{(bi)})}{\sum_t \exp(u_t^{(bi)\mathrm{T}}u^{(bi)})} \tag{8}$$

$$v^{(bi)} = \sum_t \alpha_t^{(bi)} h_t^{(bi)} \tag{9}$$

In the above equations, we first feed the bilingual word annotations $h_t^{(bi)}$ through a one-layer perceptron to get $u_t^{(bi)}$ as a hidden representation of $h_t^{(bi)}$ (Eq. 7), and then measure the importance of the word by measuring the similarity of $u_t^{(bi)}$ with a word-level context vector $u^{(bi)}$ obtaining a normalized importance weight α_t (Eq. 8). After that, we compute the bilingual attention vector v^{bi} as a weighted sum of the word annotations based on the weight (Eq. 9). The context vector $u^{(bi)}$ can be seen as a high-level informative representation of the words in memory networks (Kumar et al., 2015). The word context vector $u^{(bi)}$ are randomly initialized and jointly learned during the training process.

Monolingual Attention. To reward to the most relevant words from the two monolingual contexts for emotion classification, we again use an attention functions on the monolingual context to measure their importance.

$$u_t^{(mo)} = \tanh(W^{(mo)} h_t^{(mo)} + b^{(mo)}) \tag{10}$$

$$\alpha_t^{(mo)} = \frac{\exp(u_t^{(mo)\mathrm{T}} u^{(mo)})}{\sum_t \exp(u_t^{(mo)\mathrm{T}} u^{(mo)})} \tag{11}$$

$$v^{(mo)} = \sum_t \alpha_t^{(mo)} h_t^{(mo)} \tag{12}$$

Since we only consider monolingual contexts, the input of attention function is the annotation h_t of Chinese or English words respectively. When Chinese is used as the monolingual context, $v^{(cn)}$ equals $v^{(mo)}$ in Eq. 12. and $v^{(cn)}$ equals $v^{(mo)}$ when considering English as the monolingual context.

3.3 Prediction

The monolingual vector ($v^{(cn)}$, $v^{(en)}$) and bilingual vector ($v^{(bi)}$) with the attention mechanism mentioned above are concatenated into a single vector $F = [v^{(cn)}, v^{(en)}, v^{(bi)}]$. We then use a softmax classifier to predict the label y give the inputs X. The classifier takes the feature vector $f \in F$ as input:

$$\widehat{p}_\theta(y|X) = \mathrm{softmax}(W^{(s)} f + b^{(s)}) \tag{13}$$

$$\widehat{y} = \arg\max_y \widehat{p}_\theta(y|X) \tag{14}$$

Training cost function is the negative log-likelihood of the true class labels $y^{(k)}$ at each labeled node ($k \in \{0, 1\}$):

$$J(\theta) = -\frac{1}{2} \sum_{k=0}^{1} \log \widehat{p}_\theta(y^{(k)}|X^{(k)}) + \frac{\lambda}{2} ||\theta||_2^2 \tag{15}$$

where the superscript k indicates the k^{th} labeled node, and λ is an L2 regularization hyperparameter.

We apply online training, where model parameters are optimized by using Adagrad (Duchi et al., 2011). In order to avoid over-fitting, dropout (Hinton et al., 2012) is used to the word embedding with a ratio of 0.2. For LSTM models, we empirically set size of the hidden layer is 32. We train the word embedding using the *Skip-gram* algorithm[2].

4 Experiments

4.1 Dataset and Statistics

We focus on emotion prediction in code-switching text which is defined as text that contains more than one language ("code"). We use Chinese and English code-switching posts for experimental study, and the data set is taken from Weibo.com, one of the popular Chinese social media websites. Our dataset is collected and annotated by Wang et al. (2015). Following their setting, five basic emotions are defined as candidate emotions, namely *happiness*, *sadness*, *fear*, *anger* and *surprise*. After removing those posts containing noise and advertisements, there are 3,530 posts express emotions.

[2] `https://code.google.com/p/word2vec/`

Emotion	Chinese	English	Both
Happiness	0.670	0.127	0.203
Sadness	0.835	0.046	0.119
Anger	0.706	0.068	0.226
Fear	0.901	0.026	0.073
Surprise	0.883	0.055	0.062

Table 1: Joint distribution of emotion and causal situations.

Emotions can be expressed through the two languages separately or collectively, and we focus four types of causal situations for each emotion, namely, *None*, *English*, *Chinese*, and *Both*. *None* means that the post does not contain any corresponding emotions (E5). *Chinese* means that the emotion is expressed through the Chinese text only (E6). *English* means that the emotion is expressed through the English text only (E1). *Both* means that the emotions of the post are expressed through both Chinese and English text (E2).

The joint distribution between emotions and caused situations is illustrated in Table 1. Each cell presents the conditional probability $p(l_j|e_i)$ of the emotion e_i based on the situation l_j. From the table, we can find that, most of emotional posts are expressed through Chinese text due to Chinese being the major language. Although English text contains relatively fewer words in each post, 20.1% of emotional posts are expressed through English. It indicates that English is of vital importance to emotion expression even in code-switching contexts dominated by Chinese. And more notably, 13.7% emotional posts are conducted in both Chinese and English. It indicates that Chinese and English text would be influenced by each other, and the bilingual context would also be effective for predicting emotion in code-switching text.

E 5. 还木有看《起风了》，**mark**一下，还有《虞美人盛开的山坡》。
(I've never seen the "Kaze tachinu" and "Kokuriko-zaka kara", **mark** as a note.)

E 6. 静静坐下来看别人**show**啦。刚刚在节目里看到妈咪和弟的视频真的很意外！
(I sat down quietly to watch someone else's **show**. To my surprise, both my mother and brother appeared on the programme.)

4.2 Experiment Setting

After constructing the dataset, we randomly selected half of the annotated posts as the training data and the other half as the testing data. We use FudanNLP[3] for Chinese word segmentation and adopt the F1-Measure (F1.) to evaluate the performance of emotion prediction.

4.3 Baselines

Our first group of experiments is to investigate whether our proposed approach improves emotion prediction in code-switching text compared with state-of-the-art monolingual emotion prediction methods. For fair comparison, the following models are implemented.

- *Term-Counting (TC)* counts the Chinese and English emotional cue words for each post to predict the emotion (Tunery, 2002).
- *SVM* is the basic model which uses all the Chinese and English text of each post as features[4], we consider bag-of-words as features.
- *BLP-BS* is proposed by Wang et al., (2015), employs a Bipartite graph based Label Propagation framework with lexical level Bilingual and Sentimental information. We re-implement their approach.

[3] `https://github.com/FudanNLP/fnlp`

[4] SVM^{light} is used as the implementation for the SVM classifier, `http://svmlight.joachims.org`

Emotion	**TC**	**SVM**	**BLP-BS**	**LSTM**	**BAN**
Happiness	0.258	0.591	0.638	0.662	0.678
Sadness	0.207	0.573	0.628	0.614	0.634
Anger	0.194	0.677	0.700	0.659	0.728
Fear	0.187	0.719	0.693	0.700	0.728
Surprise	0.211	0.548	0.560	0.575	0.594
Average	0.211	0.622	0.645	0.642	0.672

Table 2: Comparison with baselines.

	CN	EN	All	Comb
LSTM	0.627	0.579	0.642	0.656
Attention	0.635	0.590	0.663	0.672

Table 3: Influence of Different Factors.

- *LSTM* uses the mixed code-switching text as the input to train a basic LSTM model, using the last hidden state vector h_n directly for emotion prediction. This services as a neural network baseline without different monolingual and bilingual context.

- *BAN* is our proposed model, which uses attention to capture the informative words from both monolingual and bilingual context.

Table 2 shows the experimental results. From the table we can find that, 1) the performance of Term-counting is unacceptable since many emotions are expressed implicitly and the importance of words is highly context-dependent. 2) Since the neural network model can capture richer features automatically, LSTM outperforms the bag-of-word SVM model in most emotions. 3) Our proposed BAN model significantly outperforms all other approaches (p-value < 0.01). This indicates the effectiveness of attention mechanism from bilingual and monolingual context, compared to learning the monolingual information. Moreover, as BAN outperforms the BLP-BS model, it shows that our proposed model can automatically capture more valuable information. Note that, BLP-BS use many external resources, such as bilingual and sentimental dictionary, while our proposed model does not use any external resources.

4.4 Influence of Different Factors

Table 3 shows the influence of different factors on our proposed model. *LSTM* represents use basic LSTM as the prediction model, and *Attention* represents the LSTM model with attention. *CN* means only considering the Chinese text of each post as input, *EN* means only considering English text, *All* means using the mixed code-switching text as input. Specificlly, *LSTM-Comb* means merge the output of LSTM-CN, LSTM-EN and LSTM-All into the full connected layer, and then get the prediction results with softmax function, and *Attention-Comb* is same as the proposed BAN model that integrate both monolingual and bilingual attention mechanisms from the document representation.

The table shows us that 1) as the English is the minor language in our corpus, the results of LSTM-EN and Attention-EN are relatively weak. 2) the model using mixed code-switching text (*All*) always outperforms the model using the monolingual text individually, indicating the bilingual information is more useful than mere monolingual information. 3) Exploring monolingual and bilingual information collectively, the LSTM-Comb model can improve the accuracy over the basic LSTM model. It indicates both monolingual (i.e., Chinese and English) and bilingual texts are useful for predicting emotion in code-switching text. 4) By aggregating the representation of the informative words, the attention-based model always outperform the traditional LSTM model. 5) Our model Attention-Comb (BAN) that further utilizes attention mechanism to capture the informative words from both monolingual and bilingual context can outperforms the previous models.

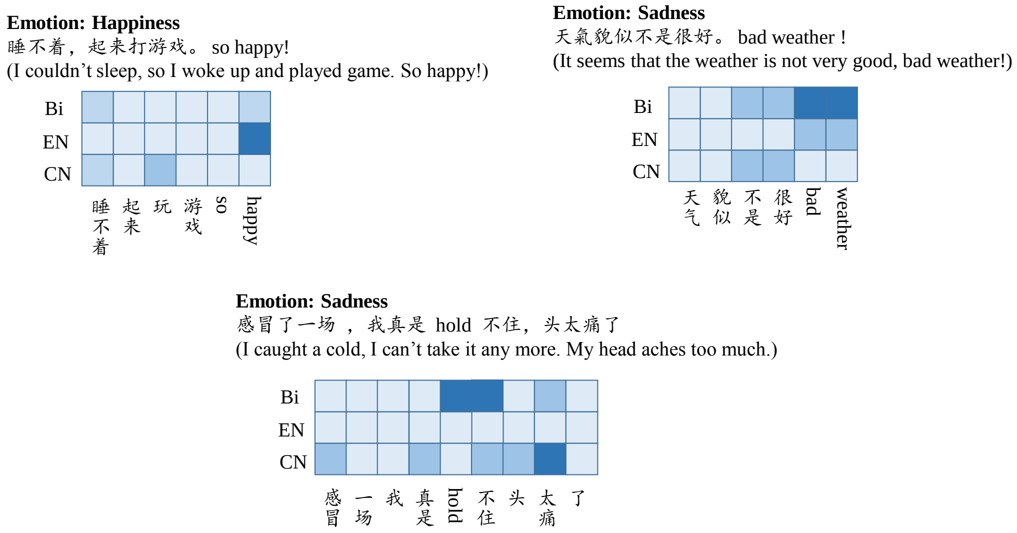

Figure 3: Example attention results.

4.5 Visualization of Attention

In order to validate that our model is able to select informative words in a post, we visualize the attention layers for several posts from our corpus in Figures 3 . Blue denotes the word weight. Since we have three attention mechanisms for monolingual and bilingual context, the first blue line denote the word weight for bilingual attention, and the other two lines denote the word weight for the Chinese and English attention respectively.

The figure shows that our model can select the words carrying strong sentiment signals such "happy", "bad", and "很好"(*very good*). In addition, since different attention functions consider different contexts, the monolingual attention functions always select the monolingual sentimental words with corresponding language such as, "happy", and "很好"(*very good*). The bilingual attention function can select mixed sentimental phrases, such as "hold不住" (*can't hold any more*). The joint attention mechanism can also deal with complex contexts. For example, in the first case, the weight of the sadness emotional word "睡不着" (*couldn't sleep*) is high with Chinese attention function, although the emotion of the whole post is happiness. However, by considering the English and bilingual attention functions, we can find the weight of word "happy" is higher than " 睡不着" (*couldn't sleep*), and it can lead us to the correct emotion.

5 Conclusion

In this paper, we addressed a novel yet important task, namely emotion detection in code-switching text. The challenges include that we need to consider both monolingual and bilingual information, and we need to extract the salient words from both monolingual and bilingual contexts. To address these challenges, a bilingual attention network model is proposed to capture the representations of monolingual and bilingual information in the code-switching text respectively. A LSTM model is used to construct the document-level representation of each post, and attention mechanism is used to capture the informative words from both monolingual and bilingual contexts. Empirical studies demonstrated that our model can significantly outperform several strong baselines.

6 Acknowledgments

Shoushan Li is the corresponding author. 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 a PolyU Faculty Research Grant (No. 1-ZVEK), 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

Heike Adel, Ngoc Thang Vu, and Tanja Schultz. 2013. Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 206–211.

Massih-Reza Amini, Cyril Goutte, and Nicolas Usunier. 2010. Combining coregularization and consensus-based self-training for multilingual text categorization. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 475–482.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 877–886.

Cícero Nogueira dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 69–78.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Wei Gao, John Blitzer, Ming Zhou, and Kam-Fai Wong. 2009. Exploiting bilingual information to improve web search. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1075–1083.

Dan Garrette, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick, and Dan Klein. 2015. Unsupervised code-switching for multilingual historical document transcription. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1036–1041.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pages 239–248.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. Ask me anything: Dynamic memory networks for natural language processing. *CoRR*, abs/1506.07285.

Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2014. Annotating events in an emotion corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 3511–3516.

Ying Li and Pascale Fung. 2012. Code-switch language model with inversion constraints for mixed language speech recognition. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 1671–1680.

Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. 2015. Sentence-level emotion classification with label and context dependence. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1045–1053.

Constantine Lignos and Mitch Marcus. 2013. Toward web-scale analysis of codeswitching. In *Proceedings of Annual Meeting of the Linguistic Society of America*.

Wang Ling, Guang Xiang, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 176–186.

Huanhuan Liu, Shoushan Li, Guodong Zhou, Chu-Ren Huang, and Peifeng Li. 2013. Joint modeling of news reader's and comment writer's emotions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 511–515.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 572–581.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *CoRR*, cs.CL/0205070.

Yanghui Rao, Xiaojun Quan, Liu Wenyin, Qing Li, and Mingliang Chen. 2012. Building word-emotion mapping dictionary for online news. In *The 1st International Workshop on Sentiment Discovery from Affective Data.*

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 973–981.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1422–1432.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1347–1353.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 235–243.

Zhongqing Wang, Sophia Yat Mei Lee, Shoushan Li, and Guodong Zhou. 2015. Emotion detection in code-switching texts via bilingual and sentimental information. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 763–768.

Bin Wei and Christopher J. Pal. 2010. Cross lingual adaptation: An experiment on sentiment classifications. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, Short Papers*, pages 258–262.

Shiyang Wen and Xiaojun Wan. 2014. Emotion classification in microblog texts using class sequential rules. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 187–193.

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

Min Yang, Dingju Zhu, and Kam-Pui Chow. 2014. A topic model for building fine-grained domain-specific emotion lexicon. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 421–426.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL 2016, 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, US*.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 612–621.