Bridging the Gap: Improve Part-of-speech Tagging for Chinese Social Media Texts with Foreign Words

Dingmin Wang[1], **Meng Fang**[2], **Yan Song**[2], **Juntao Li**[3]
[1] Tencent Cloud AI, Shenzhen, Guangdong, China
[2] Tencent AI Lab, Shenzhen, Guangdong, China
[3] Peking University, Beijing, China
{wangdimmy, moefang, mattsure}@gmail.com lijuntao@pku.edu.cn

Abstract

Multilingual speakers often switch between languages and generate enormous quantities of cross-language data. This phenomenon is more frequent observed in social media texts, where a large body of user generated data is produced every day. Such mix-lingual and informal texts lead to a challenge for part-of-speech (POS) tagging, which is one fundamental task in natural language processing. In this paper, we propose a language-agnostic POS tagger for social media texts, which is able to learn from heterogeneous data with different genre and language type. Particularly, in order to comprehensively evaluate POS tagging performance, we propose a new tagging scheme including exclusive tags for special symbols in social media texts, and a human-annotated dataset of Chinese-English mixed social media texts is also developed. Experiments on both synthetic and real datasets show the validity and effectiveness of our model on social media texts where it outperforms state-of-the-art language-specific taggers.

1 Introduction

Part-of-speech tagging is the basic step of identifying a token's functional role within a sentence and is the fundamental step in many NLP pipeline applications. It is well known that the performance of complex NLP systems is negatively affected if one of the preliminary stages is less than perfect. For example, some tagging errors may change the semantic interpretation of an entire sentence, typically due to assigning an entirely incorrect POS category to a word, for example a Plural Noun (NNS) incorrectly tagged as a Present Tense Verb (VBZ). This alteration in the semantics has a deleterious effect on all the subsequent steps in the NLP pipeline, e.g., Syntactic Parsing, Dependency Parsing, etc. Compared with formal texts, like newswire articles, the POS tagging performance in the social media texts is still far from satisfactory (Ritter et al., 2011; Gimpel et al., 2011). Most state-of-the-art POS tagging approaches are based on supervised methods, in which a large amount of annotated data is needed to train models. However, many datasets constructed for the POS tagging task are from carefully-edited newswire articles, such as PTB (Marcus et al., 1993) and CTB (Xia, 2000), which are greatly different from social media texts. The difference in domains between training data and testing data may heavily impact the performance of approaches based on supervised methods. Hence, most state-of-the-art POS taggers cannot achieve the same performance as reported on newswire domain when applied on social media texts (Owoputi et al., 2013).

However, enormous quantities of user generated content on social media are giving increasing attention as well as valuable sources for a variety of applications, such as recommendation (Jiang and Yang, 2017), disease prediction (Paul and Dredze, 2011). Yet, in such NLP tasks, one challenge is that texts from social media platforms (e.g., Twitter[1], Weibo[2]) usually contain many informal inputs, such as acronym (as soon as possible → asap), shorthand (technology → tech), out-of-vocabulary words (meeeeee → me), etc.

Another challenge is that many mix-lingual cases exist in microblogs, which occurs frequently in such informal texts. For example, according to (Zhang et al., 2014), in Weibo[7], the mixed usage of Chinese and English is one of the most popular phenomena with informal language. To illustrate

[1] http://www.twitter.com
[2] http://www.weibo.com
[3] http://nlp.stanford.edu:8080/parser/index.jsp
[4] https://github.com/fxsjy/jieba
[5] http://ictclas.nlpir.org/nlpir/
[6] Since the three POS taggers use different tag sets, so the tags are a little different.
[7] The largest Chinese social media platform.

Sent	今天帮我book一个会议室 help me book a meeting roomt today
Gold	今天/NT 帮我/AD 预定/VV 一/CD 个/M 会议室/NN
ST[3]	今天/NT 帮我/VV book一/CD 个/M 会议室/NN
Jieba[4]	今天/t 帮/v 我/r book/eng 一个/m 会议室/n
NLPIR[5]	今天/T 帮/V 我/RR book/N 一个/MQ 会议室/N

Table 1: Tagging results on an example Chinese-English Weibo by different Chinese POS taggers. Incorrect results are marked in red.[6]

such phenomenon, one example of microblogs extracted from real Weibo texts is shown in Table 1, all of which are written in Chinese with a few English words.

In this paper, we focus on the task of annotating Chinese-English social media texts from Weibo, and implement automatic part-of-speech (POS) tagging of these texts. To this end, we propose an approach to learning a POS tagger that can be applied in truly cross-language social media texts. We discuss techniques that allow us to learn a tagger given only the amount of labeled data that contains standard monolingual languages, specifically. Here, we improve the tagging performance on Weibo texts, which involves Chinese and English, by using the semantic information from different sources of labeled data. Experimental results on both synthetic and real Weibo texts confirm the effectiveness of our method. Our contributions can be concluded as follows:

- We explore to utilize multiple sources of annotated corpora to improve performance on tagging cross-lingual Weibo texts. To this end, we extends the bi-directional long short term network with adversarial training.
- For the first time, we develop a cross-lingual microblog corpus and give a quantitative evaluation for POS tagging in such microblog corpus.
- Experimental results show that our model is better than existing state-of-the-art language specific taggers.

2 Related Work

In essence, this paper is concerned with the intersection of three topics: part-of-speech tagging, processing of social media texts, and language-switching in social media texts:

Part-of-speech tagging is widely treated as a sequence labeling problem, by assigning a unique label over each sentential word (Fang and Cohn, 2016). Early studies on sequence labeling often use the models of HMM (Kupiec, 1992) and CRF (Lafferty et al., 2001) based on manually-crafted discrete features, which can suffer the feature sparsity problem and require heavy feature engineering. Recently, neural network models have been successfully applied to sequence labeling (Collobert and Weston, 2008). Among these work, the model which uses BiLSTM for feature extraction has achieved state-of-the-art performances (Huang et al., 2015), which is exploited as the baseline model in our work.

However, regarding part-of-speech tagging social media texts, the aforementioned methods are seldom used because of limited labeled data. Two most similar earlier papers are the ARK tagger (Gimpel et al., 2011) and T-Pos (Ritter et al., 2011). Both these approaches adopt clustering to handle linguistic noise, and train from a mixture of hand-annotated tweets and existing POS-labeled data. The ARK tagger reaches 92.8 % accuracy at token level but uses a coarse, customized tagset. T-Pos is based on the Penn Treebank dataset and achieves an 88.4% token tagging accuracy.

In recent years, there have been several efforts on social media text POS tagging, but almost exclusively on Twitter and mostly for English. However, it is noted that there are limited work on POS tagging cross-language texts, especially for Chinese-English texts. (Moschitti et al., 2014) reports achieving an accuracy of over 90% on English-Hindi texts and (Lascarides et al., 2009) propose a method to combine rule-based and statistically induced taggers on handling cross-language texts. However, these work on POS tagging cross-language texts can not be directly used to Chinese-English due to the great difference between languages.

3 The Model

3.1 Overview

For most of Chinese POS taggers, there are usually two kinds of ways to tag foreign words: one is to directly tag them as "foreign words", which is oversimplified; another is to give a POS tag simply based on a rule-based method, which is easy to make incorrect tags and influences the further processing for the syntactic and semantic analysis. To

Type	Input	Segmentation & Part-Of-Speech
Zh-only	我希望你去采用下我的方法	我/PN 希望/VV 你去/AD 采用/VV 下/DT 我/PN 的/DEG 方法/NN
Mixed	我希望你去follow下我的方法	我/PN 希望/VV 你去/VV follow/NN 下/LC 我/PN 的/DEG 方法/NN

Table 2: POS tagging results by Standford POS Tagger [8]. Incorrect results are marked in red.

illustrate the challenge, we take one of the state-of-the-art Chinese POS taggers (Standford Chinese POS tagger) as an example. From Table 2, we can see that when the input only contains Chinese words, the tagger can do a completely correct tagging. But, when we replace a Chinese word "采用" with its corresponding English translation "follow", we find that the tagger gives an incorrect tag "NN" to "follow". A possible reason is that Stanford Chinese POS tagger trains only on the Chinese corpus, so it is "dull" to unseen English words. However, this situation happens a lot in social media, such as Weibo.

Our goal is to train a POS tagger for Chinese social media data, which contains user-generated content and cross-language short text, specifically Chinese-English text. Because there lacks annotated Chinese social media data, we consider making use of out-of-domain (e.g., CTB (Xia, 2000)) and labeled data from other languages (e.g., PTB (Marcus et al., 1993), ARK (Gimpel et al., 2011)), which are carefully annotated and widely used in NLP-related tasks. The basic model is shown in Figure 1, with its inputs from different sources of labeled annotated data and the output being a sequence of POS tags for the given sentence. The feasibility of our method are based on the following three points:

- We use a pre-trained cross-lingual embedding, where words across two languages share same semantic space, so their semantic proximity could be correctly quantified.
- Previous work (Yang et al., 2017) has shown that *knowledge transfer* is an effective method on improving performance on a target task with few labeled training datasets. In our setting, the knowledge learned from Chinese and English datasets can be considered as a process of *knowledge transfer*, which jointly contribute to our task of tagging cross-lingual texts.
- An adversarial network is implemented to improve the share representation, aiming at achieving better tagging performance on cross-lingual texts.

Recent advances suggest that recurrent neural networks are capable of learning useful representation information for modeling problems of sequential nature (Plank et al., 2016). In this section, we describe our social media POS tagger, which is based on bidirectional long short term memory (BiLSTM). Since there is lack of annotated social media data as training data, we consider using other out-of-domain labeled data and labeled from different languages, both of which are monolingual. Instead of the common monolingual embeddings, we use cross-lingual embeddings as a bridge between different languages. Our joint model is trained based on different labeled datasets from different domains and languages. Furthermore, we improve the proposed joint model with an adversarial training scheme.

3.2 Cross-lingual Token Representation

Considering that we need to tag texts containing different languages, i.e. Chinese and English, we hope semantics-close words from different languages can have close distributed word representations. Distributed word representations are useful in NLP applications such as information retrieval, search query expansions, or representing semantics of words. A number of methods have been explored to train and apply word embeddings using continuous models for language-specific corpora. However, Chinese- and English- embeddings trained from their own language-specific corpora usually share a totally different semantic space since each language has its own vocabulary space. Therefore, although the Chinese word "政府" shares the same knowledge semantics with its English translation "government", their distance in the distributed word representation space is not close as we expect. Specially, we adopt two approaches to train a bilingual embeddings.

3.2.1 Unsupervised Training

We adopt the method proposed in (Zou et al., 2013) to achieve bilingual embeddings. First, by

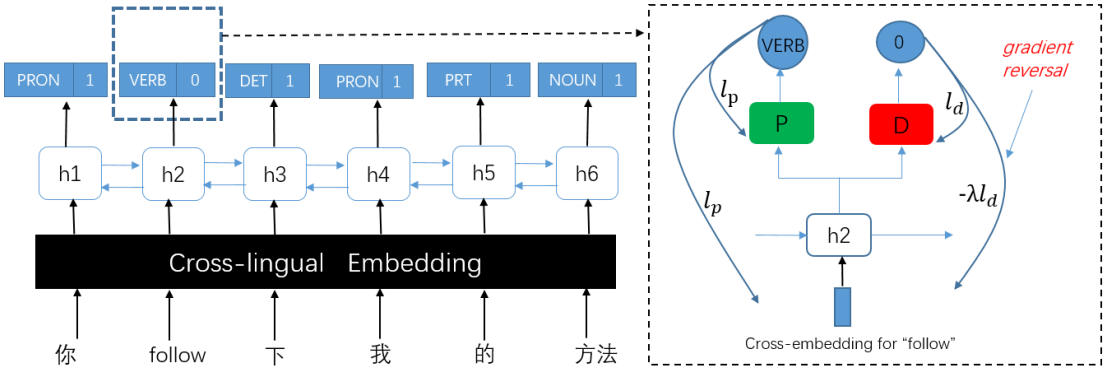

Figure 1: The general architecture of our proposed model. The green box and red box represents two feed-forward networks, which are used for the tagging task and the language identification task, respectively. Note that we only show the operation on the one hidden state of BiLSTM's outputs, and other hidden states have the same operation.

using the machine translation word alignments extracted with the Berkeley Aligner (Liang et al., 2006), two alignment matrices ($A_{zh\to en}$, $A_{en\to zh}$) are achieved. Next, two combined objectives are optimized during training:

$$J_{CO-zh} + \lambda J_{TEO-en\to zh} \quad (1)$$

$$J_{CO-en} + \lambda J_{TEO-zh\to en} \quad (2)$$

Equation 1 and 2 are optimized for Chinese embeddings and English embeddings, respectively. For example, for Chinese embedding, J_{CO-zh} is to keep the monolingual features of Chinese language itself, and $J_{TEO-en\to zh}$ is to optimize the Translation Equivalence. The embeddings are learned through curriculum training on the Chinese Gigaword corpus.

3.2.2 Embedding Projection

Instead of training embeddings joint for two languages, we consider using existing embeddings with projection. There are many available pre-trained word embeddings trained from monolingual corpora. Considering that in Weibo, most of texts are written in Chinese, we project the English embedding space (S) into the Chinese embedding space (T). Instead of re-training from parallel corpora, we adopt two methods proposed in (Song and Lee, 2017) to do the embedding projection. We get 1,000 common Chinese words and its corresponding English translations, which will be used to calculate the projection function. In linear projection, the least square fitting is used to solve the projection formula. For non-linear projection, the projection is implemented with a two-layer perceptron.

3.3 The Joint Model

To utilize a set of labeled corpora from different domains and languages to improve the tagging performance on cross-lingual Weibo texts, we first consider a joint model based on *knowledge transfer*. To facilitate this, we give an explanation for notations used in this paper. Formally, we refer to S_k as a collection of source training datasets from k labeled corpora. Mathematically,

$$S_k = \{d_i\}_{i=1}^{k} \quad (3)$$

$$d_i = \{(x_j^i, y_j^i)\}_{j=1}^{L_i} \quad (4)$$

$$x_j^i = \{w_m\}_{m=1}^{N} \quad (5)$$

$$y_j^i = \{t_m\}_{m=1}^{N}, t_m \in T, \quad (6)$$

where L_i represents the number of sentences in the corpus d_i; x_j^i and y_j^i denote a sentence and a set of tags for the sentence from d_i, respetively; N is the length of the given sentence, namely, the number of words; w_m and t_m denote a word and its corresponding POS tag, respectively; T is a set of POS tags defined in our paper, which will be described in Section 3.4.

3.3.1 The Part-of-speech Tagging Model

Let $\{x_1, \cdots, x_n\}$ be the sequence of words and $\{y_1, \cdots, y_n\}$ be the sequence of POS tags. We define the joint distribution as follows:

$$p(t_1, t_2, \cdots, t_n | x_1, x_2, \cdots, x_n) = \prod_{i=1} np(y_i | x_1, x_2, \cdots, x_n), \quad (7)$$

where $p(y_i | x_1, x_2, \cdots, x_n)$ uses a bidirectional long short term memory (BiLSTM) (Graves and

Schmidhuber, 2005). The update of each LSTM unit can be written as follows:

$$\overleftarrow{h_t}, \overrightarrow{h_t} = \text{BiLSTM}(h_{t-1}; x_t; \theta), \quad (8)$$

where x_t is a input at the current time step, h_{t-1} is hidden value of last time step, and θ represents all parameters.

For the given sequence $x = (x_1, x_2, \cdots, x_n)$, we first use an embedding layer to get the vector representation (mix-lingual embeddings) of each word x_i. The output at the last moment h_t can be regarded as the representation of the whole sequence, which has a fully connected layer followed by a softmax non-linear layer that predicts the probability distribution over classes.

$$\hat{y} = \text{softmax}(W_p(\overleftarrow{h_t} + \overrightarrow{h_t}) + b_p), \quad (9)$$

where $\hat{y}$ is prediction probabilities, W_p is the weights which need to be learned, b_p is a bias term. Given a corpus with N training samples (x_i, y_i), the parameters of the network are trained to minimise the cross-entropy of the predicted and true distributions.

$$l_p(\hat{y}, y) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j \log \hat{y}_i^j, \quad (10)$$

where y_i^j is the ground-truth label;$\hat{y}_i^j$ is prediction probabilities, and C is the class number.

3.3.2 Adversarial Training

The network described so far learns the abstract features through hidden layers that are discriminative for the part-of-speech tagging task. However, our goal is also to make these features invariant across languages in order to adapt to cross-lingual texts. To this end, we incorporate the adversarial training into our baseline POS tagger. Adversarial training (Goodfellow et al., 2014) is a powerful regularization method, which have been explored in the domain adaption (Ganin et al., 2016) and image recognition (Shrivastava et al., 2017) to improve the robustness of classifiers to input perturbations. We introduce a language discriminator, another neural network that takes the output hidden state of the BiLSTM network as input at each time step, and tries to discriminate between Chinese and English inputs in our case. Mathematically, the language discriminator is defined by a sigmoid function, and the discrimination loss is represented as the negative log-probability:

$$l_d = d \log(\hat{d}) + (1-d)\log(1-\hat{d}) \quad (11)$$

Special Word	Example	Tag
Text Emoji	:-D	EMOT
Pictorial Emoji	[:-D]	EMOJ
URLs	https://weibo.com	URL
Tel Number	88888	PHONE
At-mention	@邓超	MENT
Topic	#爸爸去哪儿#	Hash

Table 3: Six specific tags for Weibo texts

where $d \in \{0, 1\}$ denotes the language label (1 for Chinese and 0 for English), and $\hat{d}$ is the predicted probability for $d = 1$.

The overall training objective of the joint model can be written as follows:

$$l = l_p - \lambda l_d \quad (12)$$

where the hyper-parameter λ controls the relative strength of the two networks.

Specifically, in our gradient descent training, the optimization is performed by reversing the gradients of the language discrimination loss l_d (Ganin et al., 2016), when they are backpropagated to the shared layers. As shown in Figure 1, the gradient reversal is applied to the BiLSTM layer and also to the layers that come before it.

3.4 Part-of-Speech Tagsets

Since we use labeled datasets from different domains and languages, we need to map different tagsets to a uniform tagset. To do so, we use the 12 universal POS tags defined in (Petrov et al., 2011): NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories)[9].

Besides, we design additional 6 tags specific to Weibo texts: text emoticons; pictorial emoticons; URLs; telephone number; Weibo hashtags, of the form #tagname#, which the author may supply to categorize a Weibo post; and Weibo at-mentions, of the form @user, which link to other Weibo users from within a Weibo post. The details of 6 social media tags are shown in Table 3.

[9]The mapping rules for different tagsets are obtained from https://github.com/slavpetrov/universal-pos-tags.

Name	# of Sen	# of Chinese	of English	# of Other
S-weibo	1,000	10,901	1221	343
R-weibo	700	6,071	878	223

Table 4: Statistics of synthetic and manually-annotated datasets, denoted as S-weibo and M-weibo, respectively. (# of Chinese, English, Other denotes the number of words, respectively.)

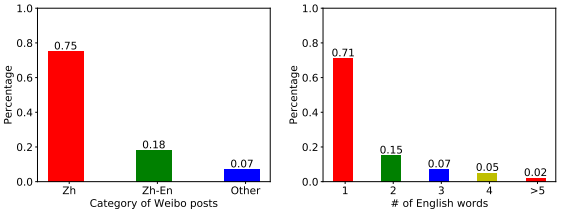

Figure 2: **Left**: Language distribution in Weibo posts. (Zh: only contains Chinese words; Zh-En: contains both Chinese and English words; Other: contains words of more than two different languages); **Right**: English words percentage in the Chinese-English mixed Weibo posts

4 Experiments

This section explains our experiments on the evaluation of our proposed model on POS tagging cross-lingual Weibo texts. First, we describe how we collect and annotate Weibo texts. A synthetic method to generate language mixed Weibo texts is also illustrated. Both of datasets are only used for testing. Next, we explore the utility of cross-lingual embeddings generated by the aforementioned two methods: unsupervised training and embedding projection. Then, we evaluate the proposed model on both the synthetic and manually-annotated datasets.

4.1 Data Collection and Annotation

Synthetic Without annotated language mixed posts from Weibo, we first propose a synthetic method to generate such data as an alternative. Considering that in Chinese-English mixed posts, English words of noun, verb and adjective categories are the most commonly used, so we randomly transform a certain percentage of Chinese words with these POS tags. An annotated Chinese-only Weibo dataset are obtained from NLPCC 2015 Shared Task (Li et al., 2015).

Manual Annotation To validate the actual performance, we develop a corpus by manually annotating text messages posted to Weibo. Initially, we collect $500,000$ raw Weibo posts using Weibo API on December 6, 2017. The posts are on various 'hot' topics (i.e., topics that are currently being discussed in news, social media, etc.). These raw posts are then divided into three categories: Chinese-only, Chinese-English and Other. The language distribution of these posts and the frequency of English words used in the Chinese-English posts shown are shown in Figure 2. We can see that over 70% mix-lingual Weibo posts only contain one English word.

Next, we randomly choose 700 posts containng both Chinese and English. Then, we ask three trained annotators to do the annotation task. Unlike English language, a Chinese word usually consists of two or more characters, so we need to segment the posts before the annotation of the POS tagging. In order to speed up the manual annotation, we first pre-segment the 700 posts using a Chinese Word Segmenter (Jieba), and these segmented posts are then proofread and modified by two trained annotators. Finally, two trained annotators are asked to tag the segmented posts using the 18 POS tags. Lastly, we ask the third annotator to tag those words that are differently tagged by the previous two annotators. Details of our experiment datasets are shown in Table 4.

4.2 Experimental Setup

Our training data mainly consists of three sources: PTB (Marcus et al., 1993), ARK (Gimpel et al., 2011), and CTB (Xia, 2000). Different tagsets are all mapped into the universal tagset described in Section 3.4. Notice that since ARK is collected from Twitter, also a kind of social media data, so we keep Tweet-specific tags and map them to our defined 6 Weibo-specific tags, for the reason that some marks also appear in Weibo texts, such as At-methion, URLs, and so on. We use three language-specific Chinese POS taggers (ST, Jieba, NLPIR) as our baseline models.

Besides, two models with and without adversarial training, denoted as BiLSTM$^-$ and BiLSTM$^+$, are implemented to study the utility of the adversarial training in our task. The hyper-parameters used for our model are as follows:

- **BiLSTM$^-$**: The hidden size is set to 150 and other hyper-parameters are tuned on a development set consisting of 10% randomly selected sentences from the training data. RMSprop (Graves, 2013) is used as optimizer.

Corpus	Models	English Word					Chinese Word	Weibo Word
		OOV	NOUN	VERB	ADJ	Other		
S-weibo	ST	\	0.611	0.802	0.557	\	0.901	0.936
	Jieba	\	0	0	0	\	0.929	0.936
	NLPI	\	0.691	0.866	0.628	\	**0.930**	0.936
	BiLSTM$^-$	\	0.701	0.863	0.708	\	0.908	0.936
	BiLSTM$^+$	\	**0.756**	**0.871**	**0.727**	\	0.912	0.936
R-weibo	ST	0.494	0.594	0.746	0.708	0.582	0.907	0.921
	Jieba	0	0	0	0	0	0.896	0.921
	NLPI	0.492	0.621	0.758	0.781	0.651	**0.918**	0.921
	BiLSTM$^-$	0.628	0.702	0.801	0.652	0.682	0.894	0.921
	BiLSTM$^+$	**0.672**	**0.731**	**0.812**	**0.703**	**0.697**	0.900	0.921

Table 5: Experimental results (F1 scores) on synthetic and manually-annoated testing datasets, denoted as **S-weibo** and **R-weibo**, respectively. In **S-weibo**, we only replace three types of English words using rules, so there is no OOV and other English words in it. Besides, the Chinese POS tagger Jieba tags all foreign words as "eng", so its F1 scores are all considered to be 0 with regard to English words.

Emebdding	Method	Train-Data	Test-data	Accuracy
Uns-emb	BiLSTM	PTB(en)	CTB(zh)	**0.511**
	BiLSTM	CTB(zh)	PTB(en)	**0.486**
Lprj-emb	BiLSTM	PTB(en)	CTB(zh)	0.346
	BiLSTM	CTB(zh)	PTB(en)	0.310
Nprj-emb	BiLSTM	PTB(en)	CTB(zh)	0.467
	BiLSTM	CTB(zh)	PTB(en)	0.406

Table 6: Experimental results by different cross-lingual embeddings (Uns-emb, Lprj-emb, Nprj-emb are cross-embeddings generated by unsupervised training, linear embedding projection and non-linear embedding projection, respectively).

- **BiLSTM$^+$**: We extend BiLSTM with an adversarial training, aiming at improving the share representation. The setting in the part of BiLSTM is same with the baseline BiLSTM. We adjust the discriminative ratio by multiple iterations of adversarial training.

4.3 Exploration of Cross-lingual Embeddings

Chinese and English have many similarities in the utterance, even if they have their own grammar rules. Therefore, with a joint semantic space of words across languages, it is possible that knowledge can be transferred from one language to another, and we can tag a corpus without having training data that has the same language with it. 6 sets of experiments are designed, where the effectiveness of cross-lingual embedding generated by three different methods is evaluated. The criteria is the POS tagging performance and we use BiLSTM as the evaluation model. In Table 6, we use cross-lingual embeddings and train a BiLSTM only on monolingual data. However, we still get a comparative cross-lingual tagging performance. In using the PTB (Marcus et al., 1993), an English annotated newswire corpus, as training data, we get a 51.1% accuracy on tagging CTB, a Chinese corpus (Xia, 2000). By the experiment, we can see that cross-lingual embeddings generated by the unsupervised training method achieve the best tagging performance. Therefore, in the following experiments, if not particularly specified, we use the cross-lingual embeddings trained by the unsupervised method.

4.4 Evaluation Results

Table 5 shows the experimental results on both synthetic and real cross-lingual Weibo posts. In terms of the tagging performance of Chinese words, our model achieves a comparable tagging performance when compared with the other three Chinese POS taggers (0.912 and 0.900, which are 0.19 and 0.18 less than the best results, respectively). One possible reason is that since our model utilizes training datasets from different languages and domains, the tag selection may be impacted by multiple factors and compromised when compared with models trained on the training data containing only one language .

However, our model achieves the best results on tagging different types of English words, which shows the effectiveness of our model. In addition, from the tagging result of Weibo words, we can see that using template rules can achieve a good performance, 0.936 and 0.921 in **S-weibo** and **R-weibo**, respectively. In Weibo (or other social

metric	sentence-level	word-level
F1-score	**0.68**	0.61

Table 7: Comparison of POS tagging performance on English words of **R-weibo** by using sentence-level and word-level translation approaches.

media texts), these social symbols are rather limited and can be easily detected, and using template rules is enough to achieve a satisfactory result.

4.5 Exploration of Translation Function

For a sentence containing both Chinese and English words, we explore the POS tagging performance by utilizing the translation system[10] and the language-specific POS tagger[11]. The language-mixed sentence is translated and then we use the language-specific POS tagger to do the tagging. In particular, we adopt two methods to do the translation as follows:

Sentence-level Translation The whole sentence is input to the translation system, and the translated results of English words may be affected by other Chinese words.

Word-level Translation In this setting, without providing the context words , we translate the English words one by one and select the first result output by the translation system if there are multiple translation results.

The experimental results on **R-weibo** are shown in Table 7. We can observe that the sentence-level translation gives the better performance. A possible reason is that with a context, the translation system can give a better translation prediction when an English word corresponds to many Chinese expressions. However, such method is over-simplified and the performance is lower than that of our proposed method shown in Table 5, which further validates the utility of our model.

Case Analysis Two real Chinese-English Weibo posts are tagged using Standford Tagger and our model, and the tagged results are shown in Table 8. In the first case, the "push" is a verb in English but is used as an adjective in the current Chinese-English text. The Standford tagger gives an incorrect tag "VERB" for "push" while our model gives the correct tagging result, which

Sent	这个老师太push＊ >⁀< ＊ Translation: This teacher is too strict ＊ >⁀< ＊
ST	这个/DET 老师/NOUN 太/ADV push/VERB ＊ >⁀< ＊/EMOJ
BiLSTM$^+$	这个/DET 老师/NOUN 太/ADV push/ADJ ＊ >⁀< ＊/EMOJ
Sent	整个场面我要Hold住 Transalation: I need to hold the whole scene
ST	整个/DET 场面/NOUN 我/PRON 要/VERB Hold/ADV 住/VERB
BiLSTM$^+$	整个/DET 场面/NOUN 我/PRON 要/VERB Hold/VERB 住/VERB

Table 8: Comparison results on two real cross-lingual Weibo posts by Standford tagger and our model, respectively. Incorrect results are marked in red.[12]

shows that our model have learned the knowledge at both syntactic and semantic level via both Chinese and English source data. Likewise, the English word "Hold" in the second case should serve as a verb as most of cases in English, but Standford tagger gives a completely incorrect tag, which indicates the constraints of language-specific taggers in handling cross-lingual texts while shows the robustness and utility of our model.

5 Conclusion

Language mixing has become a popular social phenomenon, especially in informal text such as Weibo and Twitter. In this paper, we focus on POS tagging on Chinese social media texts via learning from multiple sources of labeled corpora. To improve tagging performance on social media texts, adversarial training is adopted in our model to reduce the bias of the tagger on different languages. Experimental results confirm the validity of our approach. Compared with existing state-of-the-art language-specific taggers, our model achieves a better performance on tagging cross-lingual social media texts. We believe that our results provide a strong baseline in part-of-speech tagging Chinese social media texts.

References

Ronan Collobert and Jason Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Meng Fang and Trevor Cohn. 2016. Learning When to Trust Distant Supervision: An Application to Low-resource POS Tagging Using Cross-lingual Projection. *arXiv preprint arXiv:1607.01133*.

[10] http://fanyi.baidu.com

[11] Chinese POS tagger.http://ictclas.nlpir.org/nlpir/

[12] All tags are mapped to the twelve universal tags plus six our customized social tags as described before.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial Training of Neural Networks. *JMLR*, 17(1):2096–2030.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *ACL:short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. *arXiv pre-print*, abs/1308.0850.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv:1508.01991*.

Ling Jiang and Christopher C. Yang. 2017. User recommendation in healthcare social media by assessing user similarity in heterogeneous network. *Artificial Intelligence in Medicine*, 81:63–77.

Julian Kupiec. 1992. Robust Part-of-speech Tagging using a Hidden Markov Model. *Computer Speech & Language*, 6(3):225–242.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.

Alex Lascarides, Claire Gardent, and Joakim Nivre, editors. 2009. *EACL, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*. The Association for Computer Linguistics.

Juanzi Li, Heng Ji, Dongyan Zhao, and Yansong Feng, editors. 2015. *Natural Language Processing and Chinese Computing - 4th CCF Conference, Nanchang, China, October 9-13, 2015, Proceedings*, volume 9362 of *Lecture Notes in Computer Science*. Springer.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In *NAACL*, pages 104–111. Association for Computational Linguistics.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors. 2014. *EMNLP, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL-HLT, June, 9-14, 2013, Atlanta, Georgia, USA*, pages 380–390.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A Universal Part-of-speech Tagset. *arXiv preprint arXiv:1104.2086*.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *ACL, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *EMNLP*, pages 1524–1534. Association for Computational Linguistics.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. 2017. Learning from simulated and unsupervised images through adversarial training. In *CVPR, Honolulu, HI, USA, July 21-26, 2017*, pages 2242–2251.

Yan Song and Chia-Jung Lee. 2017. Embedding projection for query understanding. In *WWW, Perth, Australia, April 3-7, 2017*, pages 839–840.

Fei Xia. 2000. The part-of-speech tagging guidelines for the penn chinese treebank (3.0). *IRCS Technical Reports Series*, page 38.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *CoRR*, abs/1703.06345.

Qi Zhang, Huan Chen, and Xuanjing Huang. 2014. Chinese-English Mixed Text Normalization. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 433–442. ACM.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual Word Embeddings for Phrase-based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.