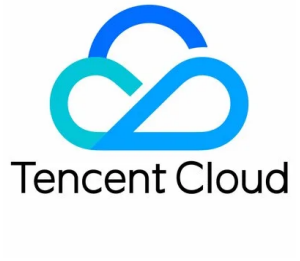
Bridging the Gap: Improve Part-of-speech Tagging for Chinese Social Media Texts with Foreign Words

Dingmin Wang[1], Meng Fang[1] , Yan Song[1] , Juntao Li[2]

[1] *Tencent*

[2]Peking University



Outline

- Introduction
- Our Work
- Experiments
- Conclusion

Outline

- **Introduction**
- Our Work
- Experiments
- Conclusion

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.
- On monolingual corpus, like PTB(Penn Treebank) and CTB(Chinese Treebank), state-of-the-art models have achieved extremely high accuracy.

The evaluation results are about **0.941** (tagging precision) in CTB test corpus of Bakeoff-4 for Chinese and **97.96** in PTB section 24 for English.

Introduction

- However, these monolingual-based models are usually trained with carefully-edited data, such as newswire articles (PTB and CTB) , whose performance will deteriorate when meeting some “foreign words” .

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]

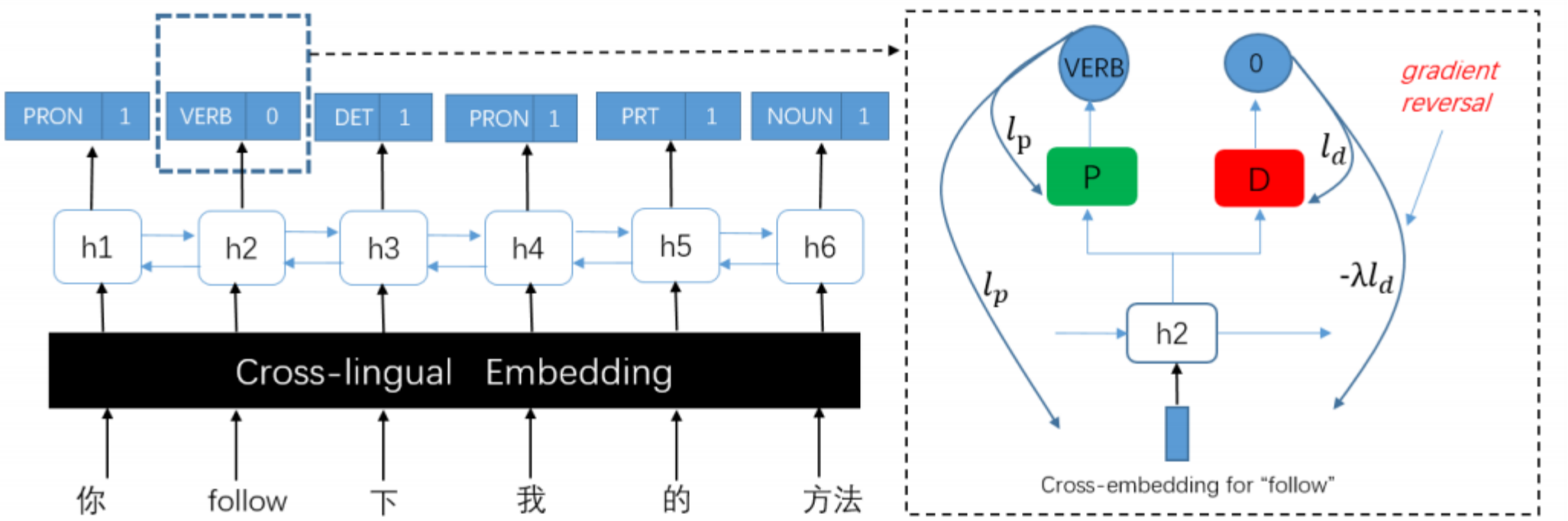
Introduction

- And the mix-lingual phenomenon is more frequent in social media texts !!!!!
- 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)
- 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.

Outline

- Introduction
- **Our Work**
- Experiments
- Conclusion

Our Work



BiLSTM with adversarial training

Our Work

- Cross-lingual Token Representation
  - **Unsupervised Training**

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

  - **Embedding Projection**

    There are many available pretrained word embeddings trained from monolingual corpora, thus we adopt two methods proposed in (Song and Lee, 2017) to do the embedding projection.

Our Work

- Joint Model

We refer to $\mathbf{S_k}$ 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.

Our Work

- 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:

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

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

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

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.

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

Our Work

- Adversarial Training

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, and the discrimination loss is represented as the negative log-probability:

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

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

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

Our Work

- Part-of-Speech Tagsets

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

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):

Besides, we design additional 6 tags specific to Weibo texts, shown in the left table.

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

Outline

- Introduction
- Our Work
- **Experiments**
- Conclusion

Experiments

- Train and test datasets

Our training data mainly consists of three sources: PTB (Marcus et al., 1993), ARK (Gimpel et al., 2011), and CTB (Xia, 2000).

The testing datasets are shown in the following table, including a synthetic one (S-weibo) and a manually-annotated one (R-weibo).

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

Experiments

Experimental results by different cross-lingual embeddings,

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

Uns-emb: unsupervised training

Lprj-emb: by linear embedding projection

Nprj-emb: non-linear embedding projection

Experiments

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

Experimental results (F1 scores) on synthetic and manually-annoated testing datasets

Experiments

- Exploration of Translation Function

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.

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

Experiments

- Case Analysis

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

Outline

- Introduction
- Our Work
- Experiments
- **Conclusion**

Conclusion

- We focus on POS tagging on Chinese social media texts via learning from multiple sources of labeled corpora, and adversarial training is adopted in our model to reduce the bias of the tagger on different languages.

- Experimental results on both synthetic and human-annotated testing datasets demonstrate the effectiveness of our model.

Thanks!

Contact: wangdimmy@gmail.com