F-Score Driven Max Margin Neural Network for Named Entity Recognition in Chinese Social Media

Hangfeng He and **Xu Sun**
MOE Key Laboratory of Computational Linguistics, Peking University
School of Electronics Engineering and Computer Science, Peking University
{hangfenghe, xusun} @pku.edu.cn

Abstract

We focus on named entity recognition (NER) for Chinese social media. With massive unlabeled text and quite limited labelled corpus, we propose a semi-supervised learning model based on B-LSTM neural network. To take advantage of traditional methods in NER such as CRF, we combine transition probability with deep learning in our model. To bridge the gap between label accuracy and F-score of NER, we construct a model which can be directly trained on F-score. When considering the instability of F-score driven method and meaningful information provided by label accuracy, we propose an integrated method to train on both F-score and label accuracy. Our integrated model yields 7.44% improvement over previous state-of-the-art result.

1 Introduction

With the development of Internet, social media plays an important role in information exchange. The natural language processing tasks on social media are more challenging which draw attention of many researchers (Li and Liu, 2015; Habib and van Keulen, 2015; Radford et al., 2015; Cherry and Guo, 2015). As the foundation of many downstream applications (Weissenborn et al., 2015; Delgado et al., 2014; Hajishirzi et al., 2013) such as information extraction, named entity recognition (NER) deserves more research in prevailing and challenging social media text. NER is a task to identify names in texts and to assign names with particular types (Sun et al., 2009; Sun, 2014; Sun et al., 2014). It is the informality of social media that discourages accuracy of NER systems. While efforts in English have narrowed the gap between social media and formal domains (Cherry and Guo, 2015), the task in Chinese remains challenging. It is caused by Chinese logographic characters which lack many clues to indicate whether a word is a name, such as capitalization. The scant labelled Chinese social media corpus makes the task more challenging (Neelakantan and Collins, 2015; Skeppstedt, 2014; Liu et al., 2015).

To address the problem, one approach is to use the lexical embeddings learnt from massive unlabeled text. To take better advantage of unlabeled text, Peng and Dredze (2015) evaluates three types of embeddings for Chinese text, and shows the effectiveness of positional character embeddings with experiments. Considering the value of word segmentation in Chinese NER, another approach is to construct an integrated model to jointly train learned representations for both predicting word segmentations and NER (Peng and Dredze, 2016).

However, the two above approaches are implemented within CRF model. We construct a semi-supervised model based on B-LSTM neural network to learn from the limited labelled corpus by using lexical information provided by massive unlabeled text. To shrink the gap between label accuracy and F-Score, we propose a method to directly train on F-Score rather than label accuracy in our model. In addition, we propose an integrated method to train on both F-Score and label accuracy. Specifically, we make contributions as follows:

- We propose a method to directly train on F-Score rather than label accuracy. In addition, we propose an integrated method to train on both F-Score and label accuracy.

- We combine transition probability into our B-LSTM based max margin neural network to form structured output in neural network.
- We evaluate two methods to use lexical embeddings from unlabeled text in neural network.

2 Model

We construct a semi-supervised model which is based on B-LSTM neural network and combine transition probability to form structured output. We propose a method to train directly on F-Score in our model. In addition, we propose an integrated method to train on both F-Score and label accuracy.

2.1 Transition Probability

B-LSTM neural network can learn from past input features and LSTM layer makes it more efficient (Hammerton, 2003; Hochreiter and Schmidhuber, 1997; Chen et al., 2015; Graves et al., 2006). However, B-LSTM cannot learn sentence level label information. Huang et al. (2015) combine CRF to use sentence level label information. We combine transition probability into our model to gain sentence level label information. To combine transition probability into B-LSTM neural network, we construct a Max Margin Neural Network (MMNN) (Pei et al., 2014) based on B-LSTM. The prediction of label in position t is given as:

$$y_t = softmax(W_{hy} * h_t + b_y) \tag{1}$$

where W_{hy} are the transformation parameters, h_t the hidden vector and b_y the bias parameter. For a input sentence $c_{[1:n]}$ with a label sequence $l_{[1:n]}$, a sentence-level score is then given as:

$$s(c_{[1:n]}, l_{[1:n]}, \theta) = \sum_{t=1}^{n} (A_{l_{t-1}l_t} + f_\Lambda(l_t|c_{[1:n]})) \tag{2}$$

where$f_\Lambda(l_t|c_{[1:n]})$ indicates the probability of label l_t at position t by the network with parameters Λ, A indicates the matrix of transition probability. In our model, $f_\Lambda(l_t|c_{[1:n]})$ is computed as:

$$f_\Lambda(l_t|c_{[1:n]}) = -log(y_t[l_t]) \tag{3}$$

We define a structured margin loss $\Delta(l, \bar{l})$ as Pei et al. (2014):

$$\Delta(l, \bar{l}) = \sum_{j=1}^{n} \kappa \mathbf{1}\{l_j \neq \bar{l}_j\} \tag{4}$$

where n is the length of setence x, κ is a discount parameter, l a given correct label sequence and $\bar{l}$ a predicted label sequence. For a given training instance (x_i, y_i), our predicted label sequence is the label sequence with highest score:

$$l_i^* = \mathop{\arg\max}_{\bar{l}_i \in Y(x_i)} s(x_i, \bar{l}_i, \theta)$$

The label sequence with the highest score can be obtained by carrying out viterbi algorithm. The regularized objective function is as follows:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} q_i(\theta) + \frac{\lambda}{2} ||\theta||^2 \tag{5}$$

$$q_i(\theta) = \max_{\bar{l}_i \in Y(x_i)} (s(x_i, \bar{l}_i, \theta) + \Delta(l_i, \bar{l}_i)) - s(x_i, l_i, \theta) \tag{6}$$

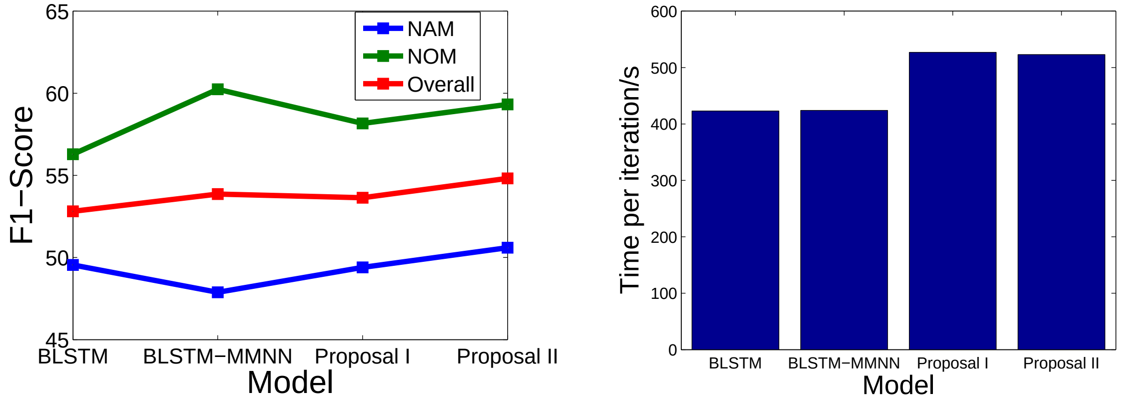
By minimizing the object, we can increase the score of correct label sequence l and decrease the score of incorrect label sequence $\bar{l}$.

2.2 F-Score Driven Training Method

Max Margin training method use structured margin loss $\Delta(l, \bar{l})$ to describe the difference between the corrected label sequence l and predicted label sequence $\bar{l}$. In fact, the structured margin loss $\Delta(l, \bar{l})$ reflect the loss in label accuracy. Considering the gap between label accuracy and F-Score in NER, we introduce a new training method to train directly on F-Score. To introduce F-Score driven training method, we need to take a look at the subgradient of equation (5):

$$\frac{\partial J}{\partial \theta} = \frac{1}{m} \sum_{i=1}^{m} (\frac{\partial s(x, \bar{l}_{max}, \theta)}{\partial \theta} - \frac{\partial s(x, l, \theta)}{\partial \theta}) + \lambda \theta \tag{7}$$

In the subgradient, we can know that structured margin loss $\Delta(l, \bar{l})$ contributes nothing to the subgradient of the regularized objective function $J(\theta)$. The margin loss $\Delta(l, \bar{l})$ serves as a trigger function to conduct the training process of B-LSTM based



(a) F-Score of the models.

(b) Running time of the models.

Figure 1: Comparing different models.

MMNN. We can introduce a new trigger function to guide the training process of neural network.

F-Score Trigger Function The main criterion of NER task is F-score. However, high label accuracy does not mean high F-score. For instance, if every named entity's last character is labeledas O, the label accuracy can be quite high, but the precision, recall and F-score are 0. We use the F-Score between corrected label sequence and predicted label sequence as trigger function, which can conduct the training process to optimize the F-Score of training examples. Our new structured margin loss can be described as:

$$\widetilde{\Delta}(l,\bar{l}) = \kappa * FScore \tag{8}$$

where $FScore$ is the F-Score between corrected label sequence and predicted label sequence.

F-Score and Label Accuracy Trigger Function The F-Score can be quite unstable in some situation. For instance, if there is no named entity in a sentence, F-Score will be always 0 regardless of the predicted label sequence. To take advantage of meaningful information provided by label accuracy, we introduce an integrated trigger function as follows:

$$\hat{\Delta}(l,\bar{l}) = \widetilde{\Delta}(l,\bar{l}) + \beta * \Delta(l,\bar{l}) \tag{9}$$

where β is a factor to adjust the weight of label accuracy and F-Score.

Because F-Score depends on the whole label sequence, we use beam search to find k label sequences with top sentece-level score $s(x,\bar{l},\theta)$ and then use trigger function to rerank the k label sequences and select the best.

2.3 Word Segmentation Representation

Word segmentation takes an important part in Chinese text processing. Both Peng and Dredze (2015) and Peng and Dredze (2016) show the value of word segmentation to Chinese NER in social media. We present two methods to use word segmentation information in neural network model.

Character and Position Embeddings To incorporate word segmentation information, we attach every character with its positional tag. This method is to distinguish the same character at different position in the word. We need to word segment the text and learn positional character embeddings from the segmented text.

Character Embeddings and Word Segmentation Features We can treat word segmentation as discrete features in neural network model. The discrete features can be easily incorporated into neural network model (Collobert et al., 2011). We use word embeddings from a LSTM pretrained on MSRA 2006 corpus to initialize the word segmentation features.

3 Experiments and Analysis

3.1 Datasets

We use the same labelled corpus[1] as Peng and Dredze (2016) for NER in Chinese social media. Details of the data are listed in Table 1. We also use

[1] We fix some labeling errors of the data.

Methods	Named Entity			Nominal Mention		
	Precision	Recall	F1	Precision	Recall	F1
Character+Segmentation	48.52	39.23	43.39	58.75	**47.96**	52.91
Character+Position	**65.87**	**39.71**	**49.55**	**68.12**	**47.96**	**56.29**

Table 2: Two methods to incorporate word segmentation information.

Models	Named Entity			Nominal Mention				
	Precision	Recall	F1	Precision	Recall	F1	Overall	OOV
(Peng and Dredze, 2015)	57.98	35.57	44.09	63.84	29.45	40.38	42.70	-
(Peng and Dredze, 2016)	63.33	39.18	48.41	58.59	37.42	45.67	47.38	-
B-LSTM	65.87	39.71	49.55	68.12	47.96	56.29	52.81	13.97
B-LSTM + MMNN	65.29	37.80	47.88	**73.53**	51.02	**60.24**	53.86	17.90
F-Score Driven I (proposal)	66.67	39.23	49.40	69.50	50.00	58.16	53.64	17.03
F-Score Driven II (proposal)	**66.93**	**40.67**	**50.60**	66.46	**53.57**	59.32	**54.82**	**20.96**

Table 3: NER results for named and nominal mentions on test data.

	Named	Nominal
Train set	957	898
Development set	153	226
Test set	209	196
Unlabeled Text	112,971,734 Weibo messages	

Table 1: Details of Weibo NER corpus.

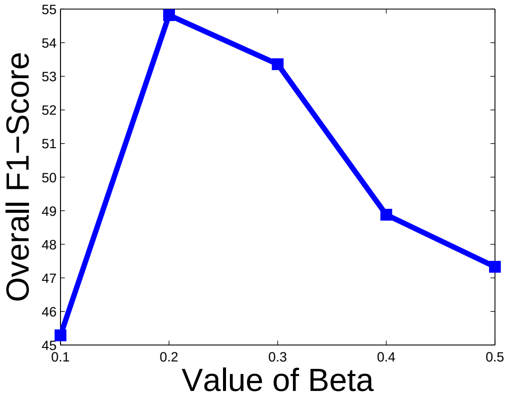


Figure 2: Overall F1-Score with different values of beta.

the same unlabelled text as Peng and Dredze (2016) from Sina Weibo service in China and the text is word segmented by a Chinese word segmentation system Jieba[2] as Peng and Dredze (2016) so that our results are more comparable to theirs.

3.2 Parameter Estimation

We pre-trained embeddings using word2vec (Mikolov et al., 2013) with the skip-gram training model, without negative sampling and other default parameter settings. Like Mao et al. (2008), we use bigram features as follow:

$$C_nC_{n+1}(n = -2, -1, 0, 1) \quad and \quad C_{-1}C_1$$

We use window approach (Collobert et al., 2011) to extract higher level Features from word feature vectors. We treat bigram features as discrete features (Collobert et al., 2011) for our neural network. Our models are trained using stochastic gradient descent with an L2 regularizer.

As for parameters in our models, window size for word embedding is 5, word embedding dimension, feature embedding dimension and hidden vector dimension are all 100, discount κ in margin loss is 0.2, and the hyper parameter for the $L2$ is 0.000001. As for learning rate, initial learning rate is 0.1 with a decay rate 0.95. For integrated model, β is 0.2. We train 20 epochs and choose the best prediction for test.

3.3 Results and Analysis

We evaluate two methods to incorporate word segmentation information. The experiment results of two methods are shown as Table 2. We can observe that positional character embeddings perform better in neural network. This is probably because positional character embeddings method can learn word segmentation information from unlabeled text while word segmentation can only adjust on training cor-

[2] https://github.com/fxsjy/jieba.

pus.

We adopt positional character embeddings in our next four models. Our first model is a B-LSTM neural network (baseline). To take advantage of traditional model (Chieu and Ng, 2003; Mccallum et al., 2001) such as CRF, we combine transition probability in our B-LSTM based MMNN. We design a F-Score driven training method in our third model F-Score Driven Model I . We propose an integrated training method in our fourth model F-Score Driven Model II .The results of models are depicted as Figure 1. From the figure, we can know our models perfrom better with little loss in time.

Table 3 shows results for NER on test sets. In the Table 3, we also show micro F1-score (Overall) and out-of-vocabulary entities (OOV) recall. Peng and Dredze. (2016) is the state-of-the-art NER system in Chinese Social media. By comparing the results of B-LSTM model and B-LSTM + MTNN model, we can know transition probability is significant for NER. Compared with B-LSTM + MMNN model, F-Score Driven Model I improves the result of named entity with a loss in nominal mention. As for the loss in nominal mention, it may be caused by the sentences without a named entity or nominal mention. The detailed analysis can be found in section 2.2. The integrated training model (F-Score Driven Model II) benefits from both label accuracy and F-Score, which achieves a new state-of-the-art NER system in Chinese social media. Our integrated model improves 2.19% on named entity and 13.65% on nominal mention.
To better understand the impact of the factor β, we show the results of our integrated model with different values of β in Figure 2. From Figure 2, we can know that β is an important factor for us to balance F-score and accuracy. Our integrated model may help alleviate the influence of noise in NER in Chinese social media.

4 Conclusions and Future Work

The results of our experiments also suggest directions for future work. We can observe all models in Table 3 achieve a much lower recall than precision (Pink et al., 2014). So we need to design some methods to solve the problem.

References

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015. Long short-term memory neural networks for chinese word segmentation. In Llus Mrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *EMNLP*, pages 1197–1206. The Association for Computational Linguistics.

Colin Cherry and Hongyu Guo. 2015. The unreasonable effectiveness of word representations for twitter named entity recognition. In *Proc. NAACL*.

Hai Leong Chieu and Hwee Tou Ng. 2003. Named entity recognition with a maximum entropy approach. In *Conference on Natural Language Learning at Hlt-Naacl*, pages 160–163.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Agustn Delgado, Raquel Martnez, Vctor Fresno, and Soto Montalvo. 2014. A data driven approach for person name disambiguation in web search results. In *COLING 2014, the International Conference on Computational Linguistics*.

Alex Graves, Santiago Fernndez, Faustino Gomez, and Jrgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference*, pages 369–376.

Mena B Habib and Maurice van Keulen. 2015. Need4tweet: a twitterbot for tweets named entity extraction and disambiguation.

Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke S. Zettlemoyer. 2013. Joint coreference resolution and named-entity linking with multi-pass sieves. In *EMNLP*, pages 289–299. ACL.

James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 172–175. Association for Computational Linguistics.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Chen Li and Yang Liu. 2015. Improving named entity recognition in tweets via detecting non-standard words. In *ACL (1)*, pages 929–938. The Association for Computer Linguistics.

Yudong Liu, Clinton Burkhart, James Hearne, and Liang Luo. 2015. Enhancing sumerian lemmatization by unsupervised named-entity recognition. In Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, editors, *HLT-NAACL*, pages 1446–1451. The Association for Computational Linguistics.

Xinnian Mao, Yuan Dong, Saike He, Sencheng Bao, and Haila Wang. 2008. Chinese word segmentation and named entity recognition based on conditional random fields. In *IJCNLP*, pages 90–93.

Andrew Mccallum, Dayne Freitag, and Fernando C. N. Pereira. 2001. Maximum entropy markov models for information extraction and segmentation. *Proc of Icml*, pages 591–598.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Arvind Neelakantan and Michael Collins. 2015. Learning dictionaries for named entity recognition using minimal supervision. *Computer Science*.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for chinese word segmentation. In *ACL (1)*, pages 293–303.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. pages 548–554.

Nanyun Peng and Mark Dredze. 2016. Learning word segmentation representations to improve named entity recognition for chinese social media. *arXiv preprint arXiv:1603.00786*.

Glen Pink, Joel Nothman, and James R. Curran. 2014. Analysing recall loss in named entity slot filling. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *EMNLP*, pages 820–830. ACL.

Will Radford, Xavier Carreras, and James Henderson. 2015. Named entity recognition with document-specific kb tag gazetteers. In Llus Mrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *EMNLP*, pages 512–517. The Association for Computational Linguistics.

Maria Skeppstedt. 2014. Enhancing medical named entity recognition with features derived from unsupervised methods. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Xu Sun, Takuya Matsuzaki, Daisuke Okanohara, and Jun'ichi Tsujii. 2009. Latent variable perceptron algorithm for structured classification. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 1236–1242.

Xu Sun, Wenjie Li, Houfeng Wang, and Qin Lu. 2014. Feature-frequency-adaptive on-line training for fast and accurate natural language processing. *Computational Linguistics*, 40(3):563–586.

Xu Sun. 2014. Structure regularization for structured prediction. In *Advances in Neural Information Processing Systems 27*, pages 2402–2410.

D. Weissenborn, L. Hennig, F. Xu, and H. Uszkoreit. 2015. Multi-objective optimization for the joint disambiguation of nouns and named entities.