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*

Overview

Motivation

Named entity recognition (NER) in Chinese social media is important but difficult because of its informality and strong noise. Although F-score is the main criterion of NER, previous work mainly train on label accuracy. To shrink the gap between label accuracy and F-Score, we propose a method to directly train on F-Score. In addition, we propose an integrated method to train on both F-Score and label accuracy.

Contributions

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.

Model

Basic Model

We construct a Max Margin Neural Network (MMNN) based on B-LSTM to combine transition probability into B-LSTM neural network.
For a input sentence $c_{[1:n]}$ with a label sequence $l_{[1:n]}$, a sentence-level score is then given as:

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

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.
We define a structured margin loss $\Delta(l, \bar{l})$:

$$\Delta(l, \bar{l}) = \sum_{j=1}^{n} \kappa \mathbf{1}\{l_j \neq \bar{l}_j\} \quad (2)$$

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.
The regularized objective function is as follows:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} q_i(\theta) + \frac{\lambda}{2}||\theta||^2 \quad (3)$$

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

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

F-Score Driven Training Method

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 labeled as O, the label accuracy can be quite high, but the precision, recall and F-score are 0. To optimize the F-Score of training examples, our new structured margin loss can be described as:

$$\widetilde{\Delta}(l, \bar{l}) = \kappa * FScore \quad (4)$$

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 following:

$$\hat{\Delta}(l, \bar{l}) = \widetilde{\Delta}(l, \bar{l}) + \beta * \Delta(l, \bar{l}) \quad (5)$$

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.

Experiments

Datasets

We use Weibo NER corpus as (Peng and Dredze, 2016) and details of the data are listed in Table 1.

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

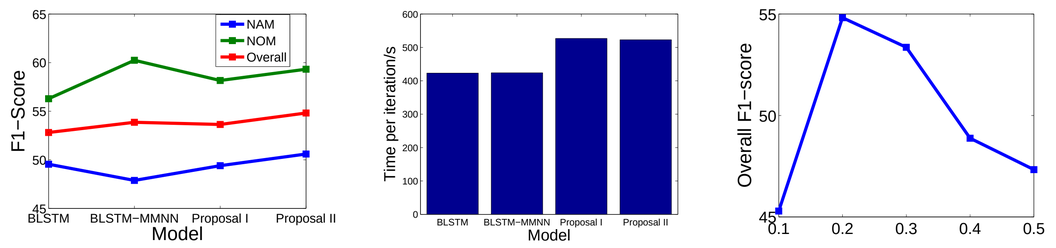
Table 1: Details of Weibo NER corpus.

Results and Analysis

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

Table 2 shows results for NER on test sets. 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. 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.



(a) F-Score of the models. (b) Running time of the models. (c) Overall F1-Score with different values of beta.

We show the results of our integrated model with different values of β in Figure 1(c). From Figure 1(c), we can know that β is an important factor for us to balance F-score and accuracy.

Conclusions

We construct a model which can be directly trained on F-score to bridge the gap between label accuracy and F-score of NER. We propose an integrated method to train on both F-score and label accuracy to make use of meaningful information provided by label accuracy. Our integrated model yields substantial improvement over previous state-of-the-art result.