Attention-based BLSTM-CRF Architecture for Mongolian Named Entity Recognition

Yuzhu Xiong
College of Computer Science
Inner Mongolia University
Huhhot 010021, China
evilbear@live.cn

Minghua Nuo*
College of Computer Science
Inner Mongolia University
Huhhot 010021, China
nuominghua@163.com

Abstract

Adding external information from unlabeled data to Named Entity Recognition task is the direction of research for solving the scarcity of labeled data. In this issue, adding pre-trained context embeddings as external information from language model is the state-of-the-art technology. However, it has combined external information to the model with simple concatenation, which assumed the same contribution of both internal and external information. While with the use of pre-trained language models, model parameters were not be updated during the iterative process of named entity recognition model. The language model cannot fully capture the semantic and syntactic roles of the contextual words in the labeled data. In view of that case, this paper proposed an improved approach by using pre-trained context embeddings based on attention mechanism. The attention mechanism layer can dynamically balance the difference between the internal information learned from bidirectional Long Short-term Memory model and external information from neural language model. Experimental results have shown that our model has achieved substantial improvement over previous one in Mongolian Named Entity Recognition task.

1 Introduction

Named entity recognition (NER) is the basic step of many natural language processing (NLP) tasks, such as Information Extraction and Machine Translation. NER's recognition precision will directly affect subsequent NLP tasks.

Mongolian has a wide range of users in China, Mongolia and Russia. However, it is a low-resource language, Mongolian NLP resources are very rare. The research on Mongolian NLP is at its initial stage. Researchers used Conditional Random Field (CRF) to predict Mongolian NER label and only did initial work on it. In terms of the NER task, we need to label the sentences with named entity tags. For example, Figure 1 gives a sentence with organization name (ORG) and person name (PER).

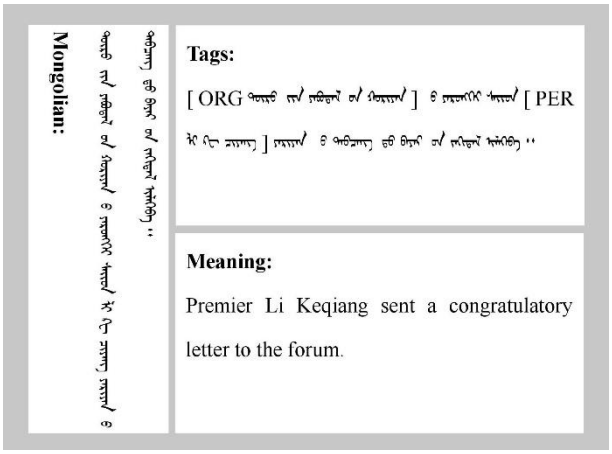

Figure 1: Example of traditional Mongolian script and named entity tags

In Mongolian corpus, data sparseness problem is more serious, due to the lexical features of Mongolian and homomorphic characters with different pronunciation, words look same but pronunciation is different in the corpus. Sometimes, it occurs to spell mistake in Mongolian corpus which is not consistent with the coding rule due to the

reason of keyboard operators' dialect. For example, the words "oyun" and "uyun" are expressed in Latin Mongolian, and they are both the word "wisdom" in English. They are the same word rather than a traditional synonym. To some extent, this phenomenon can lead to data sparsity. For further details, we have made word frequency statistics on web corpus, and there are a large number of low frequency words, some of which are misspelling of high frequency words. They have the same context, and the neural LM can learn this representation of the context sensitive and is effective for supervised NER sequence annotation tasks.

In general, neural network model can be learned by training and continuous learning on large-scale corpus. Unfortunately, it is still in the shortage of Mongolian text resource especially public labelled dataset with high quality and large-scale. So, we explored an alternate semi-supervised approach which does not require additional labeled data, it is an effective way to learn information from a large number of unlabeled dataset.

The pre-trained word embeddings (Mikolov et al., 2013; Pennington et al., 2014) from unlabeled dataset has become a fundamental component of the neural network architecture of NER tasks. Peters (2017) proposed a new approach in English NER using the unlabeled dataset pre-training language model (LM) to get context embeddings as extended information. They incorporated LM into NER model by concatenation. Because the LM embeddings are used to compute the probability of future words in a neural LM, they can capture information such as the semantic and syntactic roles of words in context.

Our research found that the way to use LM embeddings by simple concatenation is suboptimal, because it implies that the words context in the labeled dataset and the large number of unlabeled dataset are equivalent. In this case, the more similar the text style, the more effective LM embeddings will be. However, we can't get unlabeled dataset which is highly correlated with the labeled dataset. The two datasets used in this paper are news corpora from the Mongolian news website in recent years, but their editing styles are inconsistent for different programs, due to the differences between websites styles and news time spans. The pre-trained LM can’t represent the context-embedding information of the labeled dataset.

In this paper, we investigated language model extensions based on attention mechanism. In the architecture of NER model, LM concatenation layer is replaced by an attention mechanism layer. By using an attention mechanism, the model is able to dynamically balance how much information will be used between the two inputs. We named this architecture LM-ATT model. Our experiments in Mongolian shows that this architecture provides a substantial improvement over the previous model.

2 Related work

Most of the previous research on NER task in Mongolian was focused on traditional machine learning model. In recent years, researchers mainly have been using CRF (Lafferty et al., 2001; Sutton and McCallum, 2012; Wang et al., 2015, 2016a) to recognize named entity. However, CRF needs to manually set features and requires a large number of professional domain knowledge (such as digital dictionary of location names), manually, it has poor generalization ability in new areas. Wang (2016b) firstly applied the neural network model to solve the Mongolian NER task, using the BLSTM-CRF architecture. The baseline system described in this paper is consistent with Wang’s work.

Regardless of language differences, as far as NER tasks are concerned, Collobert (2011) have used the neural network model earlier, using the convolutional neural network (CNN). Huang (2015) and Lample (2016) used the BLSTM-CRF architecture model, and the experimental effect was comparable to CRF model based on rich features, which become state-of-the-art in NER task based on deep learning. Yang (2017) used transfer learning approach to design the neural network architecture for sequence tagging from cross-domain, cross-application, and cross-lingual transfer in low-resource situations. Peters (2017) demonstrated a general semi-supervised approach for adding pre-trained contextual embedding to NER task.

Attention mechanism is widely used in NLP, such as Machine Translation (Luong et al., 2015), Document Classification (Yang et al., 2016) or Sentiment Classification (Cheng et al., 2017). For NER task, Bharadwaj (2016) added phonological features to baseline neural network model and used attention mechanism on character-level to focus on more effective characters. Rei (2016) used attention

mechanism to improve the concatenation of word-level and character-level component.

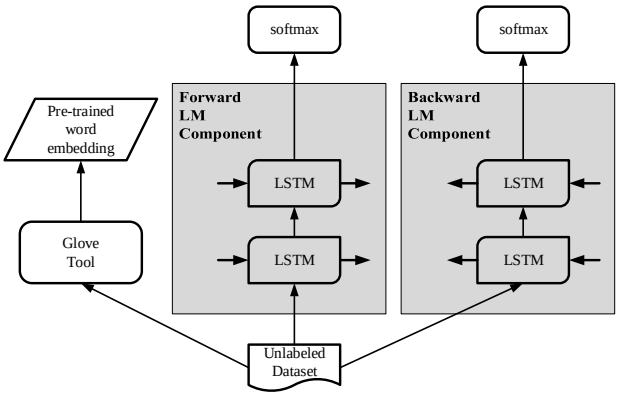

Figure 2: Overview of unlabeled dataset architecture[1]

3 Model

In this paper, our baseline system is the neural network architecture which combined BLSTM (bidirectional long short-term memory) network with CRF layer to jointly decode Mongolian NER. Then we proposed a new architecture called LM-ATT model that uses the LM embeddings as additional inputs to the sequence tagging model with attention mechanism into the baseline system. From the perspective of the data source which was used, the LM-ATT model is divided into unlabeled dataset section and labeled dataset section.

In the unlabeled dataset sub-architecture, in Figure 2, we use Glove tools[2] to get the pre-trained word embeddings, meanwhile, by constructing and pre-training neural LM to get forward and backward LM components separately. Pre-trained word embeddings and LM components from this sub-architecture are used as external information to labeled dataset sub-architecture.

Labeled dataset sub-architecture consists of three modules, as shown in Figure 3, and each of them are framed separately. It is noted that forward and backward LM components in LM representation module are pre-trained from unlabeled dataset. We remove the top layer softmax and use the forward and backward LM embeddings as input to the sequence representation module. LM components

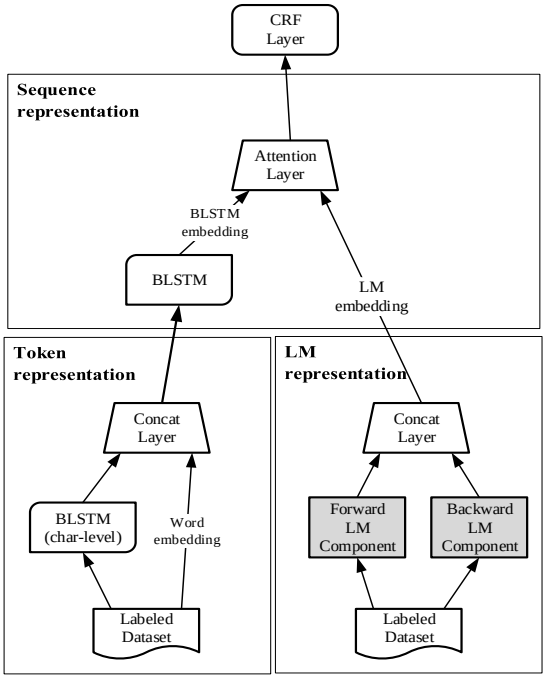

Figure 3: Overview of labeled dataset architecture[3]

only use pre-trained parameters and will not be updated in the iterations. For a Mongolian sentence to be processed, firstly we use two LM components to learn the context embedding and concatenate them to form bidirectional LM embeddings. The input of char-level BLSTM is the vectorized representation of the characters, it is randomly initialized; Word embedding is from unlabeled dataset architecture in Figure 2, which is pre-trained by Glove tool, they will be updated during the model iteration. We concatenate the characters embedding learned from char-level BLSTM with the word embedding as token representation. In order to obtain the final sentence representation for the CRF label prediction, we put token representation through BLSTM to get BLSTM representation and use LM embeddings as additional inputs to the sequence tagging model by attention-like mechanism that weights all LM embeddings in a sentence before including them in the sequence model.

In this section, we will detail the three key components of our architecture, the baseline neural network model, LM component and attention mechanism layer.

[1] The input to the LM Component is the vectorization of words in the text, which is initialized by Pre-trained word embedding.

[2] https://github.com/stanfordnlp/GloVe

[3] The LM component is the same as in Figure 2. The Concat Layer is a vector stitching.

3.1 Baseline neural network model

Our baseline neural network model (in Figure 4) is a combination of BLSTM and CRF, which is widely used in the NER task, description of BLSTM-CRF in Wang (2016b) and Lample (2016) is follows.

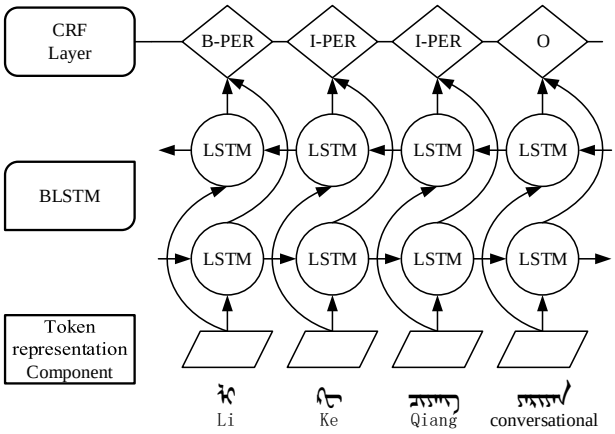

Figure 4: The main architecture of our baseline neural network

Given a sentence of words (T_1, …, T_n) from labeled dataset, all words must be represented by vector. In Mongolian corpus, token representation is the combination of characters embedding and word embedding. As shown in Figure 5, the Characters embedding is learned by the char-level BLSTM. Word embedding is a query return from Word Embedding table, which is extension of pre-trained word embedding that will be updated during model iterations. The word T_k at the k-position in the sentence is expressed as follows:

$$X_k = \left[C_k ; W_k \right] \quad (1)$$

where C_k is the Characters embedding of the word, W_k is the word embedding, X_k is the token representation of current word.

Each word in the sentence will be processed into token representation. As shown in Figure 4, BLSTM scans the words in sentence in the two directions and learns the contextual representation. The input to the BLSTM contains a hidden layer state at the previous moment in addition to the token representation. Each word is computed as follows:

$$\overrightarrow{H_k} = \overrightarrow{LSTM}\left(X_k, \overrightarrow{H_{k-1}} \right) \quad (2)$$

$$\overleftarrow{H_k} = \overleftarrow{LSTM}\left(X_k, \overleftarrow{H_{k+1}} \right) \quad (3)$$

$$H_k = \left[\overrightarrow{H_k} ; \overleftarrow{H_k} \right] \quad (4)$$

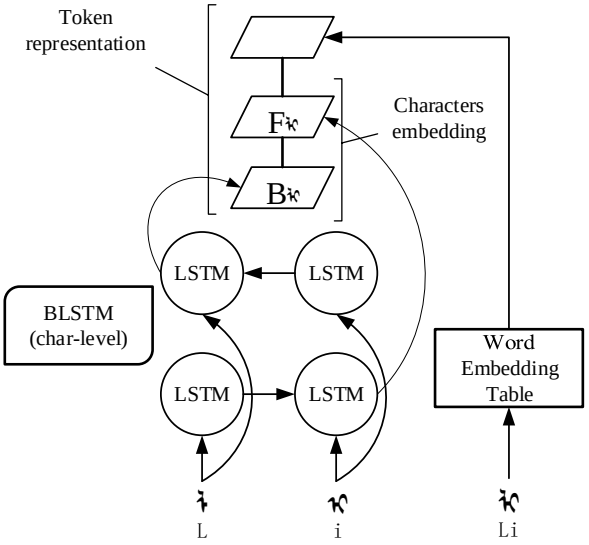

Figure 5: Details of the Token representation Component. "F" and "B" indicate the forward and backward LSTM layer separately.

where H_k is the current state of the word in BLSTM, and the arrows on $\overrightarrow{H_k}$ or $\overleftarrow{\text{LSTM}}$ indicate whether it is in the forward or backward direction.

Finally, in order to predict the label corresponding to each word in the sentence, we use CRF to decode the contextual representation. We include an extra narrow hidden layer before the CRF layer, enables the model to detect higher-level feature combinations. Since there are dependencies between tags (e.g. using the BIO labeling scheme, I-ORG can't follow B-PER), considering that this information is helpful for tag prediction, we use CRF for joint decoding. Then use the Viterbi algorithm to find the most likely tag sequence.

3.2 Recurrent Neural Networks LM

Language model is predicting the probability of the next word when given a word sequence. We have built bidirectional LM based on RNN, as shown in Figure 2. Combining LM components with softmax is a complete neural LM. It is similar to the standard LSTM-based LM described by Jozefowicz (2016). When words in sentence are vector represented, use cascade LSTM to learn forward information and then use softmax layer to predict the next word. LM uses chained rules to learn joint probabilities on word sequences:

$$p\left(T_1, ..., T_n \right) = \prod_{k=1}^{n} p\left(T_k \mid T_1, ..., T_{k-1} \right) \quad (5)$$

We also trained the reverse LM to consider the future context, in addition to the normal forward LM. Its network architecture is consistent with the forward direction, only the input, the sentence is inverted to enter, is different.

In NER task, we remove softmax layer (in Figure 2) from pre-trained forward and backward LM, rename the rest part as LM component. It should be noted that the LM embedding learned by cascade LSTMs is directly used without decoding. Because it can be combined into the NER model as a vector, and it is used to compute the probability of future word in a neural LM. Hence it can capture information such as the semantic and syntactic roles of word in context, which is very effective in NER model. Concatenate the forward and backward LM embeddings to form bidirectional LM embeddings, i.e.,

$$M_k = \left[\overrightarrow{M}_k ; \overleftarrow{M}_k\right] \quad (6)$$

where M_k is the bidirectional LM representation of a word, $\overrightarrow{M}_k$ or $\overleftarrow{M}_k$ is the one-way LM representation.

3.3 Attention mechanism layer

The embedding learned from BLSTM and LM, have various degrees of contribution to tag prediction. However, concatenation fails to consider the imbalance between them; therefore, we aim to improve by using the attention mechanism (Rei et al., 2016).

We use an attention layer to weigh the two different inputs, as shown in Figure 6. The weight parameter can control the model to dynamically determine how much each information will contribute to tag prediction. $\tanh(\ )$ and $\sigma(\ )$ are used respectively to map a weighted sum. Finally, each value of the weight matrix z is in the range [0,1]. z and Y_k represented as follows:

$$z = \sigma\left(W_z^{(1)} \tanh\left(W_z^{(2)} H_k + W_z^{(3)} M_k\right)\right) \quad (7)$$

$$Y_k = z \cdot H_k + (1 - z) \cdot M_k \quad (8)$$

where $W_z^{(1)}$, $W_z^{(2)}$ and $W_z^{(3)}$ are two-dimensional weight matrices for calculating z, they are initially randomly initialized to [-0.1, 0.1] and then automatically updated during the iteration. The vector z has the same size as x or m and is used as a weight between the two vectors. It allows the model to dynamically determine how much information to use from H_k or M_k.

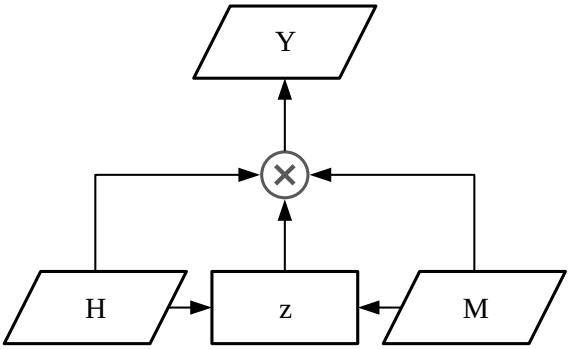

Figure 6: Internal structure of neural attention mechanism layer

4 Experiments

4.1 Datasets

We use two types of Mongolian datasets, labeled and unlabeled, and Table 1 describes some details. In data preprocessing, we put a space in front of Mongolian Unicode Character U+202F (NARROW NO-BREAK SPACE, a narrow form of a no-break space), to divide the suffix into a single word. This kind of suffix is used to express grammatical meaning, and segmentation does not affect the semantics of words, while it can alleviate the sparseness of Mongolian data to some extent.

Details	**Labeled Dataset**	**Unlabeled Dataset**
Size(MB)	24.9	362.1
Sentence number	31000	756853
PER number	9749	-
LOC number	23984	-
ORG number	15723	-

Table 1: The statistics and distribution of dataset

Labeled dataset: The labelled dataset is manually marked NER dataset from Wang (2016b). After further proofreading, this paper uses 31000 sentences, which has been marked with three kinds of named entities, that is, person name, location name and organization name. Using the BIO tags, each word has been attached one of the seven tags: B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG and O. We randomly split the labelled data into

training data (80%), development set (10%) and test set (10%).

Unlabeled dataset: The unlabeled dataset is mainly collected from most popular Mongolian news websites, and pre-processed by filtering, noise-elimination and coding conversion. All unlabeled corpus is eventually unified into Unicode encoding, approximately 362.1M.

4.2 Experiment settings

Pre-trained word embedding is obtained from the unlabeled dataset by using the Glove tool. The parameter settings for the Glove tool are described as follows. The threshold of the minimum number is 10, the context window is 15 and the word embedding dimension of each word is 300. For NER labeled dataset, word frequency statistics has been done, and all words occurs more than twice are used to develop a vocabulary and corresponding word embedding table.

Three additional words: "NUM", "UNK", "EOS" are used to build the vocabulary. All numbers in Mongolian corpus are replaced with special word "NUM", "UNK" means unregistered words, "EOS" is added to the end of the sentence in LM to distinguish different sentences. They are randomly initialized in the word embedding table. In our model architecture, each word can be represented by a word embedding stemming from word embedding table.

For the LM part, the forward and backward LMs are independently trained, but the initial parameters are consistent. They all use two layers of LSTM to learn LM embeddings to predict the next word. The main parameters are shown in Table 2.

parameter	**Baseline**	**LM**
Word LSTM layer	300	300
Char LSTM layer	150	-
Top layer	CRF	Softmax
Update embedding	True	True
Algorithm	Adam	Adam
Dropout	0.5	0.5
Epoch	100	20

Table 2: Details of model parameters

Baseline system uses mainstream experience parameters and try to fine tune. Although we set the 100 batches training period, training will stop if performance had not improved for 10 epochs. Usually, the experiment will fit around 40 iterations. The detailed parameters are shown in Table 2.

4.3 Results

We conducted experiments on the Mongolian dataset, evaluated the results by CONLL (Nadeau et al., 2007) metrics of F1. For each experiment with different architectures, we reported the average and standard deviation of 5 successful trials.

Firstly, we added the bidirectional LM component to the baseline system (BLSTM-CRF), which concatenates backward and forward LM, between BLSTM and CRF layer. In this scenario, the overall F1 score increased from 84.67 to 85.06 (see Table 3), which indicated that the external information from LM was helpful for the NER task. However, the effect is not significant enough. On analysis, the main reason was that using pre-trained LM, the information learned from LM does not represent the statement of current word well, so it is not quite effective when used directly for label prediction. Then we added an extra BLSTM layer after splicing the LM component, in order to learn further information, which follows the same architecture proposed by Peter (2017). But it did not show the same advantages as English, brought a 0.11 boost to the overall F1 score, which is not effective when applied directly to Mongolian.

Model	**F1±std**
Baseline	84.67±0.1
+LMs	85.06±0.3
+LMs+BLSTM (Peter 2017)	85.17±0.4
+LMs+ATT	**85.53**±0.3
+LMs+ATT+BlSTM	85.23±0.3

Table 3: Performance of models under different architectures

Then we combined LM by using the attention mechanism, which increases the overall F1 score by 0.47 compared with the simple concatenation. It indicates that the attention mechanism is superior, which is more suitable for the combination of inequivalent information. This is our best performing model with the overall F1 score of 85.53, we choose this model and named it as LM-ATT model. Similarly, we also add a BLSTM layer afterwards, the result indicates that newly added BLSTM played a negative role, additional layers don't work well.

Finally, we also studied the case of by using only one-way LM; results were reported in Table 4. The attention mechanism layer needs to receive the two sets of word embedding with the same dimension, but there is 600 dimensions by using the BLSTM and 300 dimensions by using the unidirectional LM, so we proposed a method to make dimension of BLSTM align with dimension of unidirectional LM, which make dimension of unidirectional LM link itself. We also found adding backward-only LM embedding outperforms forward-only LM embedding, but the gap is not large. While, the attention mechanism has brought some improvements.

Model	**F1±std**
Baseline	84.67±0.1
+LM(fw)	85.03±0.3
+LM(bw)	85.07±0.3
+LM(fw)+ATT	85.33±0.3
+LM(bw)+ATT	85.48±0.3

Table 4: Performance of model with one-way LM

In summary, attempts to add LM components in concatenation manner have improved performance relative to baseline system, but there is still room for improvement. We considered that complex morphological structure of Mongolian and simple pre-trained LM lead to the result. Architecture with the attention mechanism layer got a better performance.

5 Conclusion

In this paper, we investigate the combination of pre-trained LM and BLSTM neural network with attention for Mongolian NER task. Since the sentence information is captured by LM differs from itself, the attention mechanism performs better than concatenation when combined with the baseline system. It can dynamically balance the difference between internal and external information for NER architecture. The experimental results show that the LM-ATT model has got better performance on Mongolian NER task.

In our future work, we will consider learning external information from part-of-speech tagging (POS) task as well as LM component and using attention mechanism to combine baseline NER model with above-mentioned external information. In addition, joint training (Zheng et al., 2017) among different tasks is also worth researching.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61303165, 61773224), and Ethnic Affairs Commission of Inner Mongolia (No. MW-2018-MGYWXXH-113).

References

Bharadwaj Akash, Mortensen David, Dyer Chris, and Carbonell Jaime G. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1462–1472. Taxas, ACL.

Cheng Jiajun, Zhao Shenglin, Zhang Jiani, King Irwin, Zhang Xin, and Wang Hui. 2017. Aspect-level sentiment classification with heat (hierarchical attention) network. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pages 97–106. New York, ACM.

Collobert Ronan, Weston Jason, Bottou Léon, Karlen Michael, Kavukcuoglu Koray, and Kuksa Pavel. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12(8):2493–2537.

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

Jozefowicz Rafal, Vinyals Oriol, Schuster Mike, Shazeer Noam, and Wu Yonghui. 2016. Exploring the Limits of Language Modeling. arXiv preprint arXiv:1602.02410.

Lafferty John, McCallum Andrew, and Pereira C.N. Fernando. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, pages 282–289. San Francisco, Morgan Kaufmann.

Lample Guilaume, Ballesteros Miguel, Subramanian Sandeep, Kawakami Kazuya, and Dyer Chris. 2016. Neural Architectures for Named Entity Recognition. arXiv preprint arXiv:1603.01360.

Luong Minh-Thang, Pham Hieu, and Christoper D.Manning. 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Mikolov Tomas, Chen Kai, Corrado Greg, and Dean Jeffrey. 2013. Efficient Estimation of Word

Representations in Vector Space. arXiv preprint arXiv:1301.3781.

Nadeau David, and Sekine Satoshi. 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1):3–26.

Pennington Jeffrey, Socher Richard, and Manning Christopher. 2014. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 conference on empirical methods in natural language processing, pages:1532–1543. Doha, ACL.

Peters Matthew E., Ammar Waleed, Bhagavatula Chandra, and Power Russell. 2017. Semi-supervised sequence tagging with bidirectional language models. arXiv preprint arXiv:1705.00108.

Rei Marek, Crichton Gamal K.O., and Pyysalo Sampo. 2016. Attending to Characters in Neural Sequence Labeling Models. arXiv preprint arXiv:1611.04361.

Sutton Charles, and McCallum Andrew. 2012. An Introduction to Conditional Random Fields. Foundations and Trends® in Machine Learning, 4(4):267–373.

Wang Weihua, Bao Feilong, and Gao Guanglai. 2015. Mongolian named entity recognition using suffixes segmentation. International Conference on Asian Language Processing, pages:169–172. Suzhou, IEEE.

Wang Weihua, Bao Feilong, and Gao Guanglai. 2016a. Mongolian Named Entity Recognition System with Rich Features. The 26th International Conference on Computational Linguistics, pages:505–512. Osaka, ICCL.

Wang Weihua, Bao Feilong, and Gao Guanglai. 2016b. Mongolian Named Entity Recognition with Bidirectional Recurrent Neural Networks. 2016 IEEE 28th International Conference on Tools with Artificial Intelligence, pages:495–500. San Jose, IEEE.

Yang Zhilin, Salakhutdinov Ruslan, and Cohen William W. 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. arXiv preprint arXiv:1703.06345.

Yang Zichao, Yang Diyi, Dyer Chris, He Xiaodong, Smola Alex, and Hovy Eduard. 2016. Hierarchical Attention Networks for Document Classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages:1480–1489. San Diego, ACL.

Zheng Suncong, Hao Yuexing, Lu Dongyuan, Bao Hongyun, Xu Jiaming, Hao Hongwei, and Xu Bo. 2017. Joint entity and relation extraction based on a hybrid neural network. Neurocomputing, 257:59–66.