An Attention-Based Approach for Mongolian News Named Entity Recognition

Mingyan Tan, Feilong Bao✉, Guanglai Gao and Weihua Wang

College of Computer Science,Inner Mongolian Key Laboratory of Mongolian Information Processing Technology,Inner Mongolia University, Hohhot, China,
csfeilong@imu.edu.cn

Abstract: In the field of Natural Language Processing(NLP) of Mongolian, Named Entity Recognition(NER) has great significance. The traditional model is to use the Conditional Random Field (CRF) and Long-Short Term Model (LSTM) method. According to the characteristics of Mongolian, a named entity recognition method based on attention mechanism is proposed in this paper. According to the characteristic of the word-building of the Mongolian language, the suffix of the partial word is divided into morphemes. Based on morphemes, character vectors are trained by LSTM. After that, the word vector is sent to another LSTM to get its context representation. Then the attention mechanism is used to obtain the representation of the full text range of the character vector. Finally, the label sequence of the article is obtained by using CRF. The experimental results show that the Mongolian Named Entity Recognition of attention mechanism is superior to the traditional Bi-LSTM-CRF joint model.

Keywords: Named Entity Recognition, Attention Mechanism, Conditional Random Field, Long-Short Term Model.

1 Introduction

Named entity recognition is of great significance in the field of natural language processing. The study of named entity recognition dates back to the sixth message Understanding Conference(MUC-6)[1]. The main research topic of MUC conference is to identify and classify important nouns as well as numerical expressions such as quantity, currency, time and so on from unstructured texts. The purpose of this conference is to understand the relationship between text content mining and text extraction. With the public release of evaluation tasks on named entity recognition by conll2002[2], conll2003[3] and ace2004[4], named entity recognition has ushered in a new upsurge.

At present, the research on Mongolian named entity recognition is still relatively few. The reason is that Mongolian corpus is scarce and there is no public Mongolian corpus. Tonglaga of the Minzu University of China[5] analyzes the characteristics of the times, regional characteristics and the changing law of the internal model of human names. Seven groups of different feature templates were constructed by using conditional random field model of JingJing Cai[6] in Inner Mongolia University, and

the experiment of name recognition was carried out. On the basis of analyzing the composition characteristics of Mongolian place names, JinXing Wu[6] and others realized the named entity recognition of Mongolian place names by combining conditional random fields with dictionaries. According to the characteristics of Mongolian language, Weihua Wang [8]combines character vector and language model to realize Mongolian named entity recognition.

As a characteristic of the national minority in our country, the Mongolian language has a far-reaching significance for its research. At present, the lack of the Mongolian language, the disunity of the coding format, the derivation of the new name, different from the writing characteristics of English and Chinese, the way of word formation and so on, have brought difficulties to the identification of the Mongolian named entity. In the following, the paper will introduce the Mongolian word-formation analysis, and then cut the suffix to get the morpheme.

2 Method

This chapter first introduces the morpheme segmentation based on the characteristics of Mongolian word formation, and then introduces the Mongolian character vector. At the end of this section, we will discuss the specific algorithm and work of BiLSTM-CRF and attention mechanism in Mongolian.

2.1 Mongolian morpheme segmentation

Mongolian writing has its own characteristics: according to its position, the form of expression of a word is also different. Mongolian letters change in form at the beginning, end, and word. Therefore, in order to better show the language characteristics, this paper uses Latin form to deal with Mongolian. The contrast between Latin characters and Mongolian letters is shown in Table 1.

Mongolian suffixes can be divided into three types: word formation suffix, configuration suffix and ending suffix. The positions of the three types are relatively fixed. In a Mongolian word, there are one or more word-formation suffixes and configuration suffixes, and only one suffix at the end. However, Mongolian words can have two ending suffixes when they are added with a reverse collar suffix. If you splice all the roots with different suffixes, you can build nearly a million words. In the above three types of suffixes, the ending suffix only represents its grammatical rules and does not change the meaning of the word. Therefore, in the course of training, the ending suffix brings difficulties to the training of CRF. Figure 1 shows the relationship between root, stem, and suffix. Take the word "herdsman" as an example, the red font word "" means cow, with a green suffix "  ", it means herdsmen. The black ending suffix only represents the part of speech, and the Mongolian word means "herdsmen's."

Table 1 Comparison between Latin alphabet and Mongolian alphabet

Mongolian alphabet	Latin letter	Mongolian alphabet	Latin letter	Mongolian alphabet	Latin letter
ᠠ	a	ᠸ	w	ᡰ	L
ᠡ	e	ᠹ	f	ᡱ	Z
ᠢ	i	ᠺ	k	ᡲ	Q
ᠤ	q	ᠻ	K	ᡳ	s
ᠥ	v	ᠼ	C	ᡴ	x
ᠦ	o	ᠽ	z	ᡵ	t
ᠧ	u	ᠾ	H	ᡶ	d
ᠨ	E	ᠿ	R	ᡷ	c
ᠩ	n	ᡀ	g	ᡸ	j
ᠪ	N	ᡁ	m	᡹	r
ᠫ	b	ᡂ	l	᡺	h
ᠬ	p	ᡃ	y		

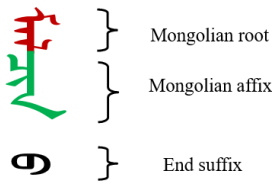

Figure 1 Character Embedding with BLSTM

2.2 Morpheme vector

According to the characteristics of Mongolian word formation, the strategy of dealing with Mongolian suffix segmentation is to divide the ending suffix of Mongolian word into a new training unit. The reason for choosing this method is that if the suffix is used as a separate training unit, the classifiers can get more context information around the suffixes to help the classifiers work. After the suffix of Mongolian word is segmented, the morpheme can be obtained, and the morpheme vector is used as the training unit for training. There are two training methods of morpheme vector, one is Continuous Bag Of Word(CBOW)[12] model, the other is skip-gram[11] model. In this paper, the skip-gram training morpheme vector is used. The specific training method is to set up a vocabulary with a vocabulary of 10000. After the word-dot code, as the input to the skip-gram model. The output of the model is a probability matrix.

The function of probability matrix is to predict contextual morphemes with current morphemes. Giving a morpheme string length *T*, Its model is represented as:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p\left(m_{t+j}\middle|m_t\right) \quad (1)$$

In formula (1), C represents the size of the context window, where the context window size is 8. For the conditional probability $p\left(m_{t+j}\middle|m_t\right)$ in formula (1), its simplest form is:

$$p(o|c)=\frac{\exp(u_o^T v_c)}{\sum_{m=1}^{M}\exp(u_o^T v_c)} \quad (2)$$

In formula (2), *o* is the ordinal number of the morpheme output and *c* is the ordinal number of the central morpheme. *U* is the morpheme output , *v* is the input morpheme. *M* is the total number of morphemes.

2.3 Character vector

Character vectors are different from morpheme vectors. Morpheme vector focuses on the semantics of the Mongolian word, while the character vector focuses on the spelling of Mongolian words, that is, the spelling of morphemes. Character vectors can be used to better describe the attributes of Mongolian words.

In this paper, LSTM is used to learn character vectors. However LSTM can only learn the information in a single direction of the sequence. Therefore, in order to learn all the characteristics of the current time sequence. The forward LSTM and the reverse LSTM need to be spliced together to form the BiLSTM[13].A forward LSTM, is used to forward processing morphemes; a reverse LSTM, is used to reverse process morphemes. The output $\overleftarrow{h}$ and $\vec{h}$ of lstm can be expressed as :

$$i_t=\sigma(W_{xi}x_i+W_{hc}h_{t-1}+W_{ci}c_{t-1}+b_i) \quad (3)$$

$$c_i=(1-i_t)\odot c_{t-i}+i_t\odot\tanh(W_{xc}x_t+W_{hc}h_{t-1}+b_c) \quad (4)$$

$$o_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+W_{co}c_t+b_o) \quad (5)$$

$$h_t=o_t\odot\tanh(c_t) \quad (6)$$

Where σ is the nonlinear activation function softmax, $\{W_{xi}, W_{hc}, W_{ci}, W_{xc}, W_{hc}, W_{xo}\ W_{ho}, W_{co}\}$is the parameter matrix of LSTM. $\{b_i\ b_c\ b_o\}$is a bias term of the model. In this paper, a 100-dimensional vector is randomly initiated for each character, and then the vector order and inverse order of the characters corresponding to the current word are input into BiLSTM respectively. The final output $[\vec{h};\overleftarrow{h}]$is the character vector of the current word.

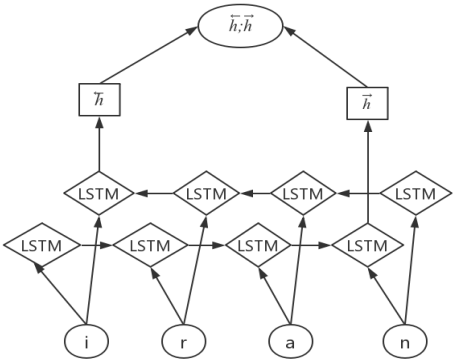

Figure 2 Character Embedding with BLSTM

2.4 BiLSTM-CRF

For a Mongolian, after the above processing, word and character vectors can be obtained. The text representation of each Mongolian word can be obtained by splicing character vector and word vector directly to BiLSTM

In the task of serial annotation, the marking strategy IOBES is widely used. In this paper, the Mongolian news corpus uses this kind of tagging form. There are three types of entities marked: "PER", "LOC" and "ORG", which represent person name, place name and organization name. In the actual annotation sequence, "I" "B" "O" tags do not appear arbitrarily, and there is a close logical relationship between them, that is the entity tags of a word are affected not only by the context of the word and the meaning of the word itself, but also by the context label of the word. However, this constraint is not taken into account in the ordinary sequence tagging model. In the label judgment of the current word, they only use the context of the current word, and do not use the context of the current word label. Therefore, in some cases, impossible tag sequences will be produced. For example, the I tag appears after the o tag and so on. In order to further improve the accuracy of entity recognition, we draw lessons from the work of Collobert et al. [9] and Huang et al. [10], combined with the advantages of crf model considering label transfer probability, this paper add the tag transfer information of the whole sentence to the original BiLSTM.

Firstly, this paper defines a label transfer matrix $\boldsymbol{A}$, where A_{ij} represents the score of the transfer from tag *i* to *j*, meanwhile parameters are trained with the model. Defines a parameter that the original BiLSTM needs to learn. Then $\theta' = \theta \cup \{A_{ij}, \forall i, j\}$ is all the parameters that the whole model needs to learn. Given a sentence $[x]_l^T$, $\boldsymbol{T}$ is the length of the sentence, Define $[f_\theta]_{it}$ the output score of the *i*th word, the *t*th label，Then the formula for calculating the total score of the first sentence in the given label sequence $[i]_l^T$ is：

$$S([x]_l^T [i]_l^T, \theta') = \sum_{t=1}^{T} (A_{[i]_{t-1},[i]_t} + [f_\theta]_{[i]_t,t}) \quad (7)$$

In this paper, we use softmax to calculate the conditional probability of a sentence $[x]_l^T$ on the real label sequence $[y]_l^T$:

$$\mathrm{p}([y]_l^T|[x]_l^T,\theta') = \frac{e^{s([x]_l^T[y]_l^T,\theta')}}{\sum_j e^{s([x]_l^T[y]_l^T,\theta')}} \tag{8}$$

In this formula, $[j]_l^T$ represents all possible label sequences. Finally, the maximum logarithmic likelihood function is used to train the model parameters. The calculation formula is:

$$\ln p([y]_l^T|[x]_l^T,\theta') = s([y]_l^T|[x]_l^T,\theta') - \ln\sum_{\forall[j]_l^T} e^{s([x]_l^T[\varphi]_l^T,\theta')} \tag{9}$$

In this paper, the random gradient drop method is used to optimize the parameters. After the training, at the end of the training, the goal is to find the tag sequence with the highest score as the prediction tag sequence, that is:

$$\mathrm{argmax}(s([y]_l^T|[x]_l^T,\theta')) \tag{10}$$

In this paper, Viterbi algorithm is used to find the best tag sequence.

2.5 Attention Mechanism

So far, one of the inevitable problems in the methods based on deep learning and traditional machine learning is the inconsistency of the full text of word labels: In one article, the same word, the same entity is often given different entity tags by the model. Obviously, this will reduce the accuracy of the model, and it is not easy to use in the actual project. The main reason for this problem is that today's models usually use sentences as separate processing units. In a separate processing unit, models assign tags according to the context of the word, that is, these models only make use of sentence information and are sentence-level methods. In the same article, if the context of the same entity in different sentences is different, the tags assigned to the sentence level model will also be different, which is the reason for the inconsistency of the full text of the word tags. At the same time, in the same article, for the same entity in its many contexts, if only one or more contexts play a decisive role in judging the label category of that entity, the current sentence-level approach does not deal with the problem well.

The sentence level method also has the problem of word label inconsistency in the specific task of Mongolian news named entity recognition. At the same time, the inconsistency of word labels is reflected in the task of organizational name entity recognition, which is the accuracy of the task. In a news article, the author usually gives the full name only when referring to an organization for the first time, and then gives it in the form of abbreviation or abbreviation. Usually, a general model can label an abbreviation correctly according to the first reference to the abbreviation. when the author refers to the entity by abbreviation, the context relationship is weak, and it is difficult for the ordinary model to assign the correct entity label according to the context information at the statement level. Therefore, in order to solve this problem, only by introducing text-level information can we better solve this problem. For this problem, this paper introduces attention mechanism to solve this problem. The attention mechanism is used to introduce text-level information, and with the help of text information, the problem is alleviated through the continuous training and learn-

ing of the model. Attention mechanism was first applied to the field of image recognition. After achieving good results in the field of image recognition, it was later applied to the field of NLP. Attention mechanism is mainly a simulated human attention mechanism [14]: When you look at an image, you don't distract your attention evenly to every part of the image, but most of them focus on specific parts of the image as needed, such as portraits, and usually focus on the face. In this paper, the attention mechanism is used to obtain text-level information for each word, and then the full-text inconsistency of the tags of the same word is improved. Specifically, for an article $[s]_l^N$, N denotes the number of sentences, x_l^T denotes one of the sentences, and t is the length of the sentence. In this paper, ***attended*** is defined as the word vector or character vector of $[s]_l^N$ and their combination; Define a one of the corresponding one of the $state_i$ for the *i*th word in the ***attended***. Define ***source*** as the context for each word in the full text, that is, the output of $[s]_l^N$ through BiLSTM. Then it can use the formula to obtain the attention α_i that the ***i***th word should allocate in the full text:

$$energy_i = f(attended, state_i, W) \tag{11}$$

$$\alpha_i = softmax(energy_i) \tag{12}$$

In formula (11), f (•) is a function used to measure the correlation between ***attended*** and $state_i$. The $\boldsymbol{W}$ in this function is trained with the model. The correlation function used in this article is Manhattan distance. Because the distance between α and oneself is 0, and the weaker the relevance of the meaning of different words, the greater the distance between them in Manhattan:

$$\mathrm{d(a, b, W)} = \sum_{t=1}^{N} w_i |a_i - b_i| \tag{13}$$

In reality, we initialize $\boldsymbol{W}$ to 1, and we keep it positive during training.
Then, we use the attention weight $\boldsymbol{\alpha}$ to filter and fuse the information in ***source*** to get the context of the current word in the full text, which we define as ***glimpse:***

$$\boldsymbol{glimpse_i} = \alpha_i^T \boldsymbol{source} \tag{14}$$

In order to make the attention model easier to train, and the entity tag of the current word depends not only on the context information within the scope of the full text, but also on the context information adjacent to the current word, this paper combines $\boldsymbol{glimpse_i}$ and inputs it into the subsequent model structure:

$$\boldsymbol{context_i} = g(\boldsymbol{glimpse}_i, \boldsymbol{source}_i, \boldsymbol{U}) \tag{15}$$

In formula(15), g (•) is a nonlinear function tanh(),$\boldsymbol{U}$ is used as a parameter trained with the model.

By using the attention mechanism and the BiLSTM, described earlier for each article $[s]_l^N$, we can get the $\sum_N \sum_T context$ (in figure 3, abbreviated as $\boldsymbol{C}$), and then through the tanh layer, it can get the score of the model on each label category for each word of the document, which is denoted as $\sum_N \sum_T output$ (abbreviated as $\boldsymbol{O}$ in Fig. 3). Finally, the total score of the article $[s]_l^N$ under the given tag sequence $\sum_M [m]_l^T$ can be calculated:

$$S([s]_l^n, \sum_M [m]_l^T, \theta') = \sum_M \sum_{t=1}^{T} (A_{[m]_{t-1}, m_t} + [output]_{[m]_t, t}) \tag{16}$$

Then, as in the previous section, the softmax function is used to obtain the probability, and the parameters of the model are trained by maximizing the logarithmic

likelihood probability. In the prediction stage, different from the previous section, Viterbi decoding is used for each sentence.

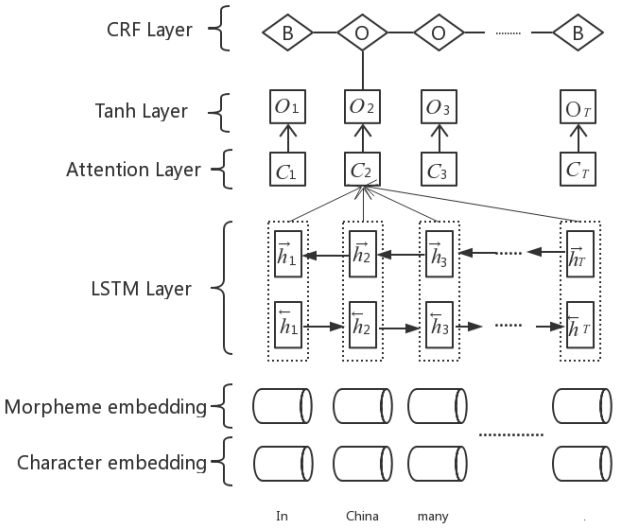

Fig.3 The model architecture of Attended-BiLSTM-CRF

3 Experimental

3.1 Experimental Setting

The traditional Mongolian language materials used in this paper come from China Mongolian News Network, people's Network (Mongolian version), Chinese Mongolian Broadcasting Network and other websites from 2013 to 2014. Through the process of mark coding correction, 33292 sentence tagging corpus is obtained, which contains 59562 entities. There are three types of entities that are marked: person name(PER), location name(LOC), organization name（ORG） Table 2 shows the number and proportion of the three entities：

Table 2 Entity type table

entity type	Number	proportion
PER	12354	20.74%
LOC	28361	47.62%
ORG	189847	31.64%

In this corpus, 10% are randomly selected as test sets and the rest as development sets. In the evaluation results, this paper uses three commonly used indicators in se-

quence tagging: the precision (p), recall rate (r), F value as the experimental evaluation index. Table 3 shows the parameters used in this model.

Table 3 The Hyper-Parameters of Model

Parameters	Description	Value
word_embeding_dim	The dimension of word embedding layer	300
Char_embedding_dim	The dimension of char embedding layer	100
Char_for_lstm_dim	The dimension of forward char LSTM layer	100
Char_rev_lstm_dim	The dimension of reverse char LSTM layer	100
For_lstm_dim	The dimension of forward LSTM layer	300
Rev_lstm_dim	The dimension of reverse LSTM layer	300
Learming_rate	Learning rate	0.001

3.2 Result

This article compares the traditional BiLSTM-CRF with the approach of joining the attention mechanism. To explore the effect of Mongolian voxel vector and common word vector on model performance. The experiment is treated in three forms：1）Which Mongolian word vector or Mongolian morpheme vector can improve the performance of the model.2) Whether the word vector or morpheme vector combined with the character vector passes through the bilstm layer, that is, whether the morpheme feature or character feature is used for the final classification.3) Whether morpheme features or character features pass through the attention layer. In other word，in what way does attention align in the full text.The experimental results are shown in Table 4.

The results can be obtained from Table 4:

1） Morpheme segmentation based on Mongolian word formation grammar can really improve the performance of label classification.

2） In LSTM layer, morpheme features combined with character features are higher than word features combined with character features. In Table 4, this result is marked in blue fonts.

3) In the attention layer, the use of character features alone is higher than the use of morpheme features alone. Moreover, the simultaneous use of morpheme features and character features in the attention layer will degrade performance. In Table 4, this result is marked in black font；

Table 4 Expermental Results for the Feature of Attention

Model	Char LSTM	Morpheme LSTM	Morpheme Attention	Char Attention	**P(%)**	**R**(%)	**F**(%)
BiLSTM-CRF	√	√	×	×	**87.26**	**87.80**	**87.55**
BiLSTM-CRF	√	×	×	×	83.99	83.98	84.48
BiLSTM-CRF	×	√	×	×	84.81	86.36	85.58
Attended-BiLSTM-CRF	×	√	√	×	88.13	87.98	88.05
Attended-BiLSTM-CRF	√	×	×	√	89.58	90.47	90.08
Attended-BiLSTM-CRF	√	√	√	√	**91.24**	**90.17**	**90.67**

Note：“√”indicates that our model uses this feature.“×”indicates that our model does not use this feature.

The reasons for the above results are as follows:

1） The attention mechanism can learn the text level information in Mongolian. Thus, the consistency of the full text and the recognition rate of abbreviations can be improved.

2） When judging the label category, it mainly depends on the meaning of the word, not the meaning of the character, and the meaning of the word and the character are not mutually exclusive. In Mongolian word formation, morphemes can roughly express the meaning of a word. Therefore, in the LSTM layer, morpheme features are better than word features, and the combination of morpheme features and character features is better.

3） There are a large number of unknown words, so when using morphemes to do attention, the unlogged words will affect the generation of attention weights, resulting in improper weight distribution. There is no such problem with characters. Therefore, the single character feature of attention layer is better than the word feature. When the two are used at the same time, the model cannot completely eliminate the shortcomings of word features, so the performance of the joint use is degraded.

3.3 Performance Comparison Experiment

In order to verify the performance of Attended-BiLSTM-CRF in Mongolian, this paper compares the performance with other methods. The method compared in this paper is the result of Wang’s experiment[8] on the corpus used in this paper. The specific experimental data of each entity are given here by the author, as shown in Table 5:

Table 5 Performance comparison experimental table

Model	Entity	**P**(%)	**R**(%)	**F**(%)
BiLSTM-CRF[8]	PER	89.26	87.19	88.21
	LOC	83.57	86.78	85.15
	ORG	88.17	85.64	86.88
Attended-BiLSTM-CRF	PER	93.31	89.16	91.18
	LOC	90.45	90.10	89.82
	ORG	90.40	87.89	89.12

it is not difficult to see from the table that the performance of each entity recognition has been improved. Among them, the recognition performance of geographical names and organizational names is greatly improved. The f value of geographical names increased by 4.67, and the f value of organizational names increased by 2.24.

4 Conclusion

In this paper, the problem of the identification of the Mongolian named entity is solved by using the method of Attended-BiLSTM-CRF on the Mongolian news material. The experimental results show that this method has better results than the existing BiLSTM-CRF method based on Mongolian, which has the following reasons：

1) Based on the characteristics of Mongolian word composition, word-element segmentation is beneficial to the training of the model.
2) Low dimension, dense Morpheme vectors and character vectors have better performance than traditional machine learning, while depth models such as LSTM, can better learn high-level abstract information.
3) attention mechanism makes use of text-level information to effectively reduce the inconsistency of the full text of word tags, and at the same time, it also improves the accuracy of abbreviation recognition.

However, in the field of Mongolian named entity recognition, attention mechanism still has a lot of room for improvement. From the experiments in this paper, it can be seen that the text-level information can effectively improve the performance, especially in abbreviated entities such as place names and organizational structure names. At the same time, the research of Mongolian named entity recognition provides a good basis for the establishment of Mongolian knowledge base and Mongolian question and answer database in the future.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Nos. 61563040, 61773224); Natural Science Foundation of Inner Mongolia (Nos. 2018MS06006，2016ZD06).

References

1. Grishman R, Sundheim B. Message Understanding Conference-6: A brief history. In: Proceedings of COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. 1-12(1996).
2. Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: Proceedings of the sixth conference on Natural language learning. Association for Computational Linguistics, 121-128(2002).
3. Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics. 142-147(2003).
4. Doddington G R, Mitchell A, Przybocki M A, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In: Proceedings of LREC. 21-6(2004, 2)
5. Tongkala. Automatic Name Recognition Based on Mongolian Corpus.Doctor,Central University for Nationalities(2013).
6. Jingjing Cai. Auotmatic Mongolian Personal Recognition Based on CRF. Doctor,Inner Mongolia University(2016).
7. Jingxin Wu,LiLi,Zhenxin Yang. Research on Mongolian Place Name Recognition Based on CRF and Dictionary.Computer Engineering and Science.38(05),1046-1051(2016).
8. Weihua Wang.Research on Mongolian Named Entity Recognition.Doctor,Inner Mongolia University(2018.)
9. Collobert R, Weston J, Bottou L, et al. Natural language processing(almost) from scratch. Journal of Machine Learning Research.12(8),2493-2537(2018)
10. Huang Zhiheng, Xu Wei, Yu Kai. BIdirectional LSTM-CRF models for sequence tagging.https://arxiv.org/abs/1508.01991(2017-07-04)
11. Word2Vec (Part 1): NLP With Deep Learning with Tensorflow (Skip-gram): http://www.thushv.com/natural_language_processing/word2vec-part-1-nlp-with-deep-learning-with-tensorflow-skip-gram/ last accessed 2018/11/21.
12. Continuous Bag of Words :https://iksinc.online/tag/continuous-bag-of-words-cbow/ last accessed 2019/05/21.
13. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks.18(5):602-610(2015).
14. Mnih V , Heess N , Graves A , et al. Recurrent Models of Visual Attention. Advances in neural information processing systems(2014)
15. Svozil D, Kvasnicka V, Pospichal J. Introduction to multi-layer feed-forward neural networks. Chemometrics and intelligent laboratory systems.39(1): 43-62(1997).
16. Luong T, Socher R, Manning C. Better word representations with recursive neural networks for morphology. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning. 104-113(2003).
17. Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the web: An experimental study[J]. Artificial intelligence.165(1): 91-134(2015).
18. Weihua Wang, Feilong Bao, Guanglai Gao. Mongolian Named Entity Recognition with Bidirectional Recurrent Neural Networks.In:Processings of the 28th IEEE International Conference on Tools with Artificial Intelligence(ICTAI). 495-500(2016).