Tibetan medical named entity recognition study for Tibetan clinical electronic medical records

Cuo Zhuoma*, Jia Cairang, Duanzhu Sangjie, Zhuoma Yangmao, Zhaxi Zhuoma
The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining 810008, Qinghai, China

ABSTRACT

In current years, there are many open corpus and research results for Chinese clinical natural language processing (for short NLP) of Biomedical and Traditional Chinese Medicine. However, the research in this area for Traditional Tibetan Medicine lags behind. We know that Traditional Tibetan Medicine has its own unique set of theory system in treating and saving people. Therefore, it is imperative to speed up the research on the Tibetan Clinical Natural Language Processing. Medical named entity recognition is an important subtask of Clinical Natural Language Processing. So Tibetan medical named entity recognition is an urgent and basic research work for Traditional Tibetan Medicine. Due to the scarcity of labeled datasets, the Medical Named Entity Recognition task of Traditional Tibetan Medicine clinical text is still an unvisited researching area. In this work, we firstly manually construct a labeled dataset for this task and then explore this area with deep learning approaches by designing a Tibetan Lattice-LSTM-CRF neural network architecture. To further improve the model performance, we also incorporate both syllable and word level pre-trained representation. The final empirical results show that the proposed models could produce accuracy rate, recall rate and F1 values of 91.89%, 93.15% and 92.52%, respectively on our test set, which shows the validity of the model in the paper.

Keywords: Traditional Tibetan medicine, Tibetan medical entity recognition, Tibetan syllable, Tibetan Lattice-LSTM-CRF

1. INTRODUCTION

Medical Named Entity Recognition (for short MNER) is one of the basic and difficult duties in the Clinical Natural Language Processing (for short CNLP)[1-3]. In current years, prevalence of deep learning methods on NER has reduced the model's dependence on artificial feature engineering and existing toolkits[4-18]. Among them, the Lattice long short-term memory (for short Lattice LSTM) model has obtained good consequences in the Chinese open domain named entity recognition task[19]. Because the Lattice-LSTM network model can fully integrate the word information and the potential vocabulary information of the word, it effectively avoids word segmentation error propagation.

The Biomedical Named Entity Recognition (for short BioNER) is an important part of MER. There are many research results on BioNER[20, 21] and some research results on Tibetan named entity recognition for open domain[22, 23]. But the scarcity of Traditional Tibetan Medicine clinical text NLP resources leads to a lack of corresponding toolkits, which has greatly influenced the development process of Tibetan Clinical Natural Language Processing (for short TCNLP). In particular, the study of Tibetan Medical Named Entity Recognition (for short TMNER) for Traditional Tibetan Medicine clinical texts is still in its infancy.

In view of this, the paper aims to put the research of entity recognition into effect in the field of Traditional Tibetan Medicine. Therefore, derived from the Lattice LSTM neural network model and the characteristics of Tibetan syllables (a Tibetan syllable is equivalent to a Chinese character), a Lattice-LSTM model suitable for Tibetan medical named entity recognition is constructed. In the Tibetan texts, syllables are the basic unit of words. The Lattice-LSTM network structure can fully integrate the syllable information and the potential word information of the syllable, which effectively avoids the wrong transmission of word segmentation. The research results indicate that the accuracy rate, recall rate and F1 values of the Tibetan medical entity recognition by the method in this paper reached 91.89%, 93.15% and 92.52%, respectively. And achieves better performance.

* 2353498508@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066W
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2662023

2. LATTICE LSTM

In the field of English, the first to use deep learning on named entity recognition task was Hammerton et al.[24]. The network structure used was a unidirectional LSTM. Due to LSTM's good sequence modeling capabilities, LSTM-CRF has become one of the basic frameworks for named entity recognition[25-27]. Many methods are based on LSTM-CRF as the main framework, which incorporates various relevant features. In this paper, the LSTM-CRF is used as the main network structure and the Tibetan syllable characteristics are combined according to the English named recognition model with good recognition effect. The Syllable Lattice LSTM model is introduced before the LSTM-CRF structure. We give an example of syllable sequences and marker sequences for entity recognition based on syllables in the field of Traditional Tibetan medicine, as shown in Table 1. The syllables in the first row of the table are derived from the following Tibetan sentences པུས་ཚིགས་ན་ཞིང་འདེགས་འཇོག་དཀའ་བ (Knee pain, can't bend).

Table 1. Examples of syllable sequences and labeling sequences for entity recognition in Tibetan medicine based on syllables.

Syllable	པུས	ཚིགས	ན	ཞིང	འདེགས	འཇོག	དཀའ	བ
Label	B-S	I-S	E-S	O	B-S	I-S	I-S	E-S

Note: O represents the external marker of the entity; B represents the start marker of the entity; I represents the internal marker of the entity; E represents the end marker of the entity; And S represents the Symptom entity.

2.1 Tibetan syllable-based model

Input vector of the Tibetan syllable-based entity recognition model (Figure 1) is a syllable sequence. There is a defect in this model, that is, the word itself and word order information of Tibetan words cannot be fully utilized.

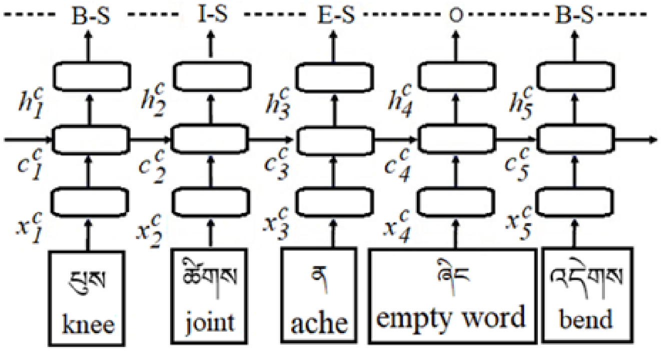

Figure 1. Tibetan syllable-based model.

2.2 Tibetan word-based model

The principle of the Tibetan word-based model (Figure 2) is similar that of the Tibetan syllable-based model, except that the input vector to this model is Tibetan word sequences after word segmentation. Therefore, this model will lead to the word segmentation error transmission. Thus it ultimately affects the performance of entity recognition.

2.3 Tibetan Lattice-LSTM-CRF model

Due to the exponential number of word-to-syllable paths in the grid, the Lattice LSTM model is employed to automatically manage the sentence's information flow from start to finish. The gating unit is used to communicate information about various paths to each syllable in a dynamic manner. Lattice LSTM can learn to automatically identify helpful phrases from the information flow after exercising on the training data set, which enhances the performance of named entity identification (Figure 3). The benefit of the approach suggested in this study over syllable-based and word-based named entity identification systems is the inclusion of explicit vocabulary information for word segmentation rather than merely automatic attention, which reduces word segmentation mistakes.

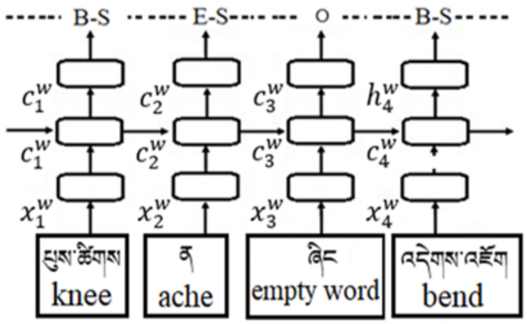

Figure 2. Tibetan word-based model.

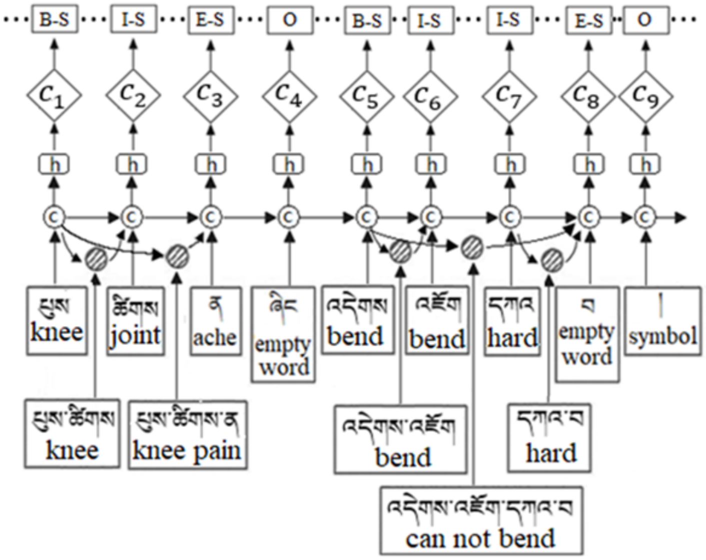

Figure 3. Tibetan Lattice-LSTM-CRF model.

2.4 LSTM layer

Although an RNN neural network may theoretically handle sequence information of any length. In practice, the gradient will disappear when the arrays are too long, and then characteristics of long-term dependency cannot be learned. Graves et al. enhanced the recurrent neural network and introduced the LSTM model to address this issue[28]. The LSTM unit regulates data transport via the input cell, forget cell, and output cell. LSTM is a special RNN that can learn long-term laws. They have been applied very well on various issues and are now widely used. The LSTM encode unit (Figure 4).

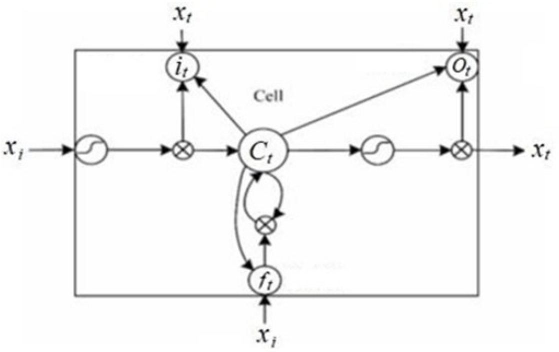

Figure 4. LSTM unit.

The specific calculation process is shown in Equations (1)-(6):

$$i_t = \sigma(W_i \cdot h_{t-1} + U_i + b_i) \tag{1}$$

$$f_t = \sigma(W_f \cdot h_{t-1} + U_f \cdot x_t + b_f) \tag{2}$$

$$\widetilde{c_t} = \tanh(W_c \cdot h_{t-1} + U_c \cdot x_t + b_f) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \tag{4}$$

$$o_t = \sigma(W_o \cdot h_{t-1} + U_o \cdot x_t + b_o) \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where σ is the sigmoid function and ⊙ is the dot product. At time t, x_t indicates the input status, h_t indicates the hidden state, and o_t indicates the output status, which comprises all effective information from the previous *t*. The control information goes into the following instant through c_t, which is an update gate. Because f_t indicates a forget gate, that control messages are erased. The output of the concealed state is determined jointly by the two.

3. EXPERIMENT AND RESULT ANALYSIS

3.1 Experiment data

3.1.1 Experiment Data. This article takes news data comprising 60,000 syllables from the Chinese Tibetan Netcom website and trains them using the Glove model to create Tibetan syllable vectors[29]. The Tibetan syllable vector formed has a dimension of 50.

3.1.2 Experiment Dataset. Since the current entity identification in the Tibetan medicine field lacks a publicly labeled data set, this paper annotates the existing 530 electronic medical records to construct an entity identification data set in the Tibetan medicine field. We randomly divided the above data in units of medical record documents, and obtained a training set of 305 documents and a test set of 225 documents. This data set contains three major categories of Tibetan medical entities: symptom (SYMPTOM), disease (DISEASE), and prescription (PRESCRIPTION). Table 2 displays the distribution of the number of categories.

Table 2. Entity recognition dataset of Tibetan medicine.

Dataset	Number of corpus	SYMPTOM(S)	DISEASE(D)	PRESCRIPTION(P)
Training set	305	3025	1980	1220
Test set	225	2231	1742	900
Total	530	5256	3722	2120

3.2 Annotation strategies and evaluation indicators

General named entity recognition and labeled strategies include BIO mode, BIEO mode, and BIEOS mode. The BIEOS labeled technique is applied in this paper, that B means the start of the entity, I means inside the entity, and E means the entity's termination. O means non-entity or entity outside, S means single syllable entity. When predicting the boundary of an entity, the entity type needs to be predicted at the same time, so there are 11 different categories of tags that can be predicted are O, S, B-S, I-S, E-S, B-D, I-D, E-D, B-P, I-P, E-P. In the testing process, it is only determined that an entity's prediction is precision when both its border and type are entirely accurate.

The identification performance evaluation indexes of entities in the field of Tibetan medicine involve precision rate (P), recall rate (R) and composite index F1. The specific calculation method is shown in Equation (7), where *Tp* is the total number of entities that the model successfully identified, *Fp* is the total number of unrelated entities that the model correctly identified, and *Fn* is the total number of related entities that the model failed to identify.

$$\mathrm{P} = \frac{T_P}{T_P + F_P} \times 100\%$$

$$R = \frac{T_P}{T_P + F_n} \times 100\% \quad (7)$$

$$F = \frac{2P \cdot R}{P + R} \times 100\%$$

3.3 Experiment environment and super parameter settings

The experiment environment in this study is Python 2.7, and the deep learning framework is Pytorch 0.3.0. post4. The settings of the neural network's hyperparameter will have an effect on the neural network's performance. Table 3 displays the settings of neural network parameters in this paper.

Table 3. Neural network hyper parameter values.

Parameter	Value	Parameter	Value
Syllable vector dimension	50	Word vector dimension	50
Lattice vector dimension	50	Lattice drop rate	0.5
Dropout	0.5	Learning rate	0.05
LSTM layer	1	Dimension of the main LSTM hidden layer	200

3.4 Experiment design and results

The three categories of entities in the Tibetan medical named entity recognition dataset were utilized to test the recognition performance of the model used in this work, the following two sets of experiments were designed. The evaluation indexes of the experiment are precision (P), recall rate (R) and comprehensive index F1 value.

Experiment 1 Comparison of Tibetan medical named entity recognition results in Traditional Tibetan Medicine clinical text by different models. Table 4 gives the recognition results of the CRF (Conditional Random Field) model and the Lattice LSTM model, respectively.

Table 4. Performance comparison of various models for Tibetan medical named entity recognition (%).

Model	P	R	F1
CRF	90.34	88.67	89.97
Tibetan Lattice-LSTM	91.89	93.15	92.52

The findings of the experiment demonstrate that the Lattice LSTM model's F1 value is 2.55% greater than the CRF model's F1 value, demonstrating that the deep learning model outperforms the model based on statistics for the entity recognition problem in the field of Tibetan medicine. And it has been demonstrated that the neural network model can decrease the model's reliance on artificial feature engineering by employing only representation features.

Experiment 2 Lattice LSTM model and CNN-BiLSTM-CRF model based on word vectors were used to contrast the effectiveness of entity recognition in the field of Tibetan medicine. The Lattice LSTM model can concurrently encode word information relating to the sequence and syllable-level sequence information for the model to access. Lattice LSTM adds word information in addition to syllable granularity, which improves the semantic expression and effectively solves the issue of improper word segmentation transfer. Table 5 shows that the Lattice LSTM model can effectively improve Tibetan medical named entity recognition effect.

Figure 5 shows the change trend of the number of iterations of the neural network and the values of the three entity recognition indexes P, R and F1.

Table 5. Comparison of performance between word vector-based model and Lattice LSTM model (%).

Model	P	R	F1
Word-CNN-BiLSTM-CRF	88.89	88.81	89.87
Tibetan Lattice-LSTM	91.89	93.15	92.52

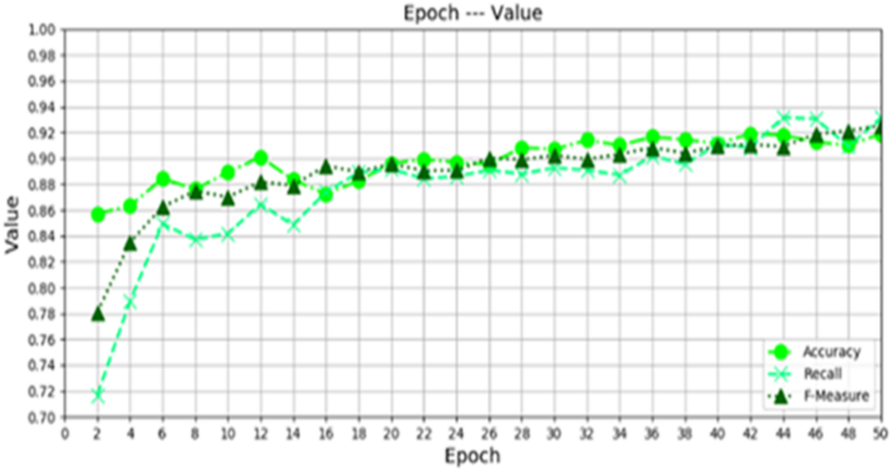

Figure 5. The number of iterations and the trend of P, R and F1 values.

Table 6 shows the accuracy rate (P), recall rate (R), and comprehensive index F1 value for the three entities SYMPTOM, DISEASE, and PRESCRIPTION.

Table 6. The recognition outcomes of Tibetan medical named entities of different types.

Model	Type	P	R	F1
Tibetan Lattice-LSTM	SYMPTOM	92.77	91.61	92.36
	DISEASE	92.01	94.91	94.34
	PRESCRIPTION	87.63	91.11	88.76

The predictive accuracy of prescription entities in the above table is low. It is mainly due to the presence of nested, acronyms and other interference information in many Tibetan drug names in prescriptions. So it is easy to predict errors without other sufficient context information. For example， the name of Traditional Tibetan medicine བསམ་འཕེལ་ནོར་བུ (Ruyi Treasure Pills) is often abbreviated to བསམ་ནོར (Meaningless) in prescriptions. In addition, because there are not many medical records collected at present, there are fewer nested disease entities involved in it. However, as the amount of data increases, the disease names become more and more nested. This condition will have an impact on the accurate identification of disease entities. For example, the both disease entities ཁྲག་འཁྲུགས་སྙིང་ལ་བབས་པའི་ནད (Hypertensive heart disease) and བད་ཀན་མེ་ཉམས་ཀྱི་ནད (Chronic superficial gastritis) have been nesting phenomena. Based on the above phenomena, we will focus on solving the problems of nesting and abbreviation in Traditional Tibetan medicine entities in our future research work.

3.5 Example of Tibetan medical named entity recognition

Take a sentence in a Tibetan inpatient medical record as an example to demonstrate the Tibetan medical named entity recognition effect of the Tibetan Lattice-LSTM model. Specific examples are shown in Table 7.

Table 7. Example of Tibetan medical named entity recognition.

Sentence	དགོང་མོ་ཨ་གར་བཅོ་ལྔ་ཁེ༡རེ་བསྟེན། (Take 1 gram Fifteen flavor agilawood powder in the evening)
Correct participle	དགོང་མོ(evening) ཨ་གར་བཅོ་ལྔ(Fifteen flavor agilawood powder) ཁེ (gram) ༡ (1) རེ (granules) བསྟེན (take)
Automatic word segmentation	དགོང་མོ (evening) ཨ་གར (agilawood) བཅོ་ལྔ (fifteen) ཁེ (gram) ༡ (1) རེ (empty word) བསྟེན (take)
Lattice participle	དགོང་མོ (evening) ཨ་གར (agilawood) བཅོ་ལྔ (fifteen) ཨ་གར་བཅོ་ལྔ (Fifteen flavor agilawood powder) ཁེ (gram) ཁེ༡ (1gram) རེ (empty word) བསྟེན (take)
Word-CNN-BiLSTM-CRF	དགོང་མོ་*ཨ་གར་*P*བཅོ་ལྔ་*Pཁེ༡རེ་བསྟེན། (Take 1 gram Fifteen flavor agilawood powder in the evening)
Tibetan Lattice-LSTM	དགོང་མོ་ཨ་གར་བཅོ་ལྔ་Pཁེ༡རེ་བསྟེན། (Take 1 gram Fifteen flavor agilawood powder in the evening)

Note: Italic words indicate incorrect entities, boldfaced word indicates correct entities.

4. CONCLUSION

This research suggests a recognition approach based on neural networks to address the issue of Tibetan medical named entity recognition. On the established entity recognition dataset of Traditional Tibetan medicine, experiments were done. The effective Tibetan medical entity recognition results will help build a more accurate Tibetan medical Knowledge Graph, which can be built a high-performance medical intelligent Question-and-Answer system for Traditional Tibetan medicine.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (NSFC No.61662061, 61063033). Qinghai Province Tibetan Information Processing and Machine Translation Key Laboratory (No.2020-ZJ-Y05), and Qinghai Normal University's 2020 university-level natural science young and middle-aged scientific research fund project (No.2020QZR017).

REFERENCES

[1] Zhao, S., Liu, T., Zhao, S., et al., "A neural multi-task learning framework to jointly model medical named entity recognition and normalization," Proceedings of the AAAI Conference on Artificial Intelligence, 33, 817-824(2019).

[2] Li, Y., Wang, X., Hui, L., et al., "Chinese clinical named entity recognition in electronic medical records: development of a lattice long short-term memory model with contextualized character representations," JMIR Medical Informatics, 8(9), e19848(2020).

[3] Wang, C., Wang, H., Zhuang, H., et al. "Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree," Journal of Biomedical Informatics, 111, 103583(2020).

[4] Zhang, H.-N. and Wu, D.-Y. "Chinese named entity recognition based on deep neural network," Journal of Chinese Information Processing, 31(4), 28-35(2017). (in Chinese)

[5] Ma, X. and Hovy E., "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF", (2016).

[6] Chiua J. P. C. and Eric, N., "Named entity recognition with bidirectional LSTM-CNNs," Transactions of the Association for Computational Linguistics, 4, 357-370(2016).

[7] Lample G., Ballesteros M., Subramanian S., et al., "Neural architectures for named entity recognition," arXiv preprint arXiv.1603.01360, (2016).

[8] Joel, N., Nicky, R., Will, R., Tara, M., Curran, J. R., "Learning multilingual named entity recognition from Wikipedia," Artificial Intelligence, (2013).

[9] dos Santos, C. N. and Guimarães, V., "Boosting named entity recognition with neural character embeddings," Computer Science, (2015).

[10] Maimaitiayifu, Wushouer, S., Palidan, M. and Yang, W., "Uyghur named entity recognition based on BiLSTM-CNN-CRF model," Computer Engineering, 44(8), 230-236(2018). (in Chinese)

[11] Wang, J., Zhang, R.-D., et al. "GRU-based named entity recognition method," Application of Computer Systems, 27(09), 18-24(2018). (in Chinese)

[12] Li, L.-H., Guo, Y.-K. "Biomedical named entity recognition based on CNN-BLSTM-CRF model," Journal of Chinese Information Processing, (2018). (in Chinese)

[13] He J. and Wang H., "Chinese named entity recognition and word segmentation based on character," Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, (2008).

[14] Liu, Z., Zhu, C. and Zhao, T., "Chinese named entity recognition with a sequence labeling approach: Based on characters, or based on words? Advanced intelligent computing theories and applications," With Aspects of Artificial Intelligence, Springer, Berlin Heidelberg, (2010).

[15] Yang, P., Yang, Z.-H., et al., "Name entity recognition of chemical drugs based on attention mechanism," Computer Research and Development, 55(07), 1548-1556(2018). (in Chinese)

[16] Li, H., Hagiwara, M., Li, Q., et al., "Comparison of the impact of word segmentation on name tagging for Chinese and Japanese," LREC, 2532-2536(2014).

[17] Chen, W., Zhang, Y. and Isahara, H., "Chinese named entity recognition with conditional random fields," Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, 118-121(2006).

[18] Dong, C., Zhang, J., Zong, C., et al., "Character-based LSTM-CRF with radical-level features for Chinese named entity recognition," Natural Language Understanding and Intelligent Applications. Springer, Cham., 239-250(2016).

[19] Zhang, Y. and Yang, J., "Chinese NER using Lattice LSTM," arXiv preprint, (2018).

[20] Zhao, S., Cai, Z., Chen, H., et al., "Adversarial training based lattice LSTM for Chinese clinical named entity recognition," Journal of Biomedical Informatics, 99, 103290(2019).

[21] Wen, G., Chen, H., Li, H., et al., "Cross domains adversarial learning for Chinese named entity recognition for online medical consultation," Journal of Biomedical Informatics, 112, 103608(2020).

[22] Jin, M., Yang, H.-H., et al., "Recognition of Tibetan named entity," Journal of Northwest University for Nationalities (Natural Science Edition), 31(03), 49-52(2010). (in Chinese)

[23] Huaque, C.-R., Jiang, W.-B., Zhao, H.-X., et al., "Tibetan Named entity recognition based on perceptron model," Computer Engineering and Applications, 50(15), 72-176(2014). (in Chinese)

[24] Hammerton, J., "Named entity recognition with long short-term memory," Conference on Natural Language Learning at Hlt-naacl, Association for Computational Linguistics, (2003).

[25] Huang, Z., Xu, W. and Yu, K., "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv.1508.01991, (2015).

[26] Yao, K. S., Peng, B. L., Zweig, G., et al., "Recurrent conditional random fields for language understanding," ICASSP, (2014).

[27] Yao, K. S., Peng, B., Zhang, Y., et al., "Spoken language understanding using long short-term memory neural networks," IEEE SLT, (2014).

[28] Graves A. "Generating Sequences with Recurrent Neural Networks". Computer Science. 1308.0850, 1-43 (2013).

[29] Jeffrey, P., Richard, S. and Christopher, D. M., "GloVe: 'Global vectors for word representation'," Empirical Methods in Natural Language Processing, (2014).