MTNER: A Corpus for Mongolian Tourism Named Entity Recognition

Xiao Cheng, Weihua Wang*, Feilong Bao, Guanglai Gao
College of Computer Science, Inner Mongolia University,
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology,
Huhhot, China, 010021
yychengxiao99@163.com, wangwh@imu.edu.cn

Abstract. Name Entity Recognition is the essential tool for machine translation. Traditional Named Entity Recognition focuses on the person, location and organization names. However, there is still a lack of data to identify travel-related named entities, especially in Mongolian. In this paper, we introduce a newly corpus for Mongolian Tourism Named Entity Recognition (MTNER), consisting of 16,000 sentences annotated with 18 entity types. We trained in-domain BERT representations with the 10GB of unannotated Mongolian corpus, and trained a NER model based on the BERT tagging model with the newly corpus. Which achieves an overall 82.09 F1 score on Mongolian Tourism Named Entity Recognition and lead to an absolute increase of +3.54 F1 score over the traditional CRF Named Entity Recognition method.

Keywords: Named Entity Recognition, Mongolian Tourism Corpus, NER model based on BERT

1 Introduction

Recently there has been significant interest in modeling human language together with the special domain, especially tourism, as more data become available on websites such as tourism websites and apps. This is an ambitious yet promising direction for scaling up language understanding to richer domains. There is no denying in saying that machine translation plays a pivotal role in this situation. Therefore, it is high time that we should stress the significance of machine translation.

Named Entity Recognition (NER) is defined as finding names in an open do-main text and classifying them among several predefined categories, such as the person, location, and organization names.[3] It not only is the fundamental task of Natural Language Processing (NLP), but also the basic work on machine translation. In addition, it is a very important step in developing other downstream NLP applications [3]. More importantly, it also plays an indispensable role in other natural language processing tasks, such as information extraction, information retrieval, knowledge map-

* Corresponding author

ping, knowledge map, question answering system and so on. Therefore, this is a very challenging problem in the field of natural language processing (NLP).

In recent years, Bidirectional Encoder Representation from Transformers (BERT) has performed extremely well in multiple tasks in the field of natural language processing. Most open source monolingual BERT models support English or Chinese, but none support Mongolian. For this purpose, we proposed a BERT pre-training language model suitable for Mongolian researchers, and trained a NER model based on the BERT tagging model by using our Mongolian tourism labeling corpus.

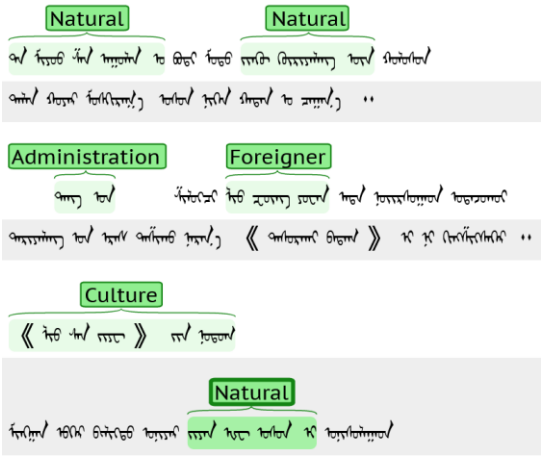


Fig. 1. Examples of Mongolian travel-related named entities in a Mongolian post.

In this paper, we present a comprehensive study to explore the unique challenges of named entity recognition in the tourism field. These named entities are often ambiguous, for example, there is one person name and one location name in the given sentence, the word "ᠬᠠᠨᠠ" commonly refers to a location name, but also be used as a personal name.

To identify these entities, we propose a NER model based on the Bert-Base tag model, which is a very powerful baseline model that identifies 18 types of travel-related named entities. This model combines local sentence-level context information with remote monitoring information. The NER model is strictly tested by using 16000 newly annotated Mongolian tourism corpus, and its performance is better than the traditional CRF model and BiLSTM-CRF model. Our major contributions are as follows:

A NER corpus for Mongolian Tourism. 18 types of named entities were manually annotated, including most Mongolian tourism information.

We demonstrate that Named Entity Recognition in the tourism domain is an ideal benchmark task for testing the effectiveness of contextual word representations, such as ELMo and BERT, due to its inherent polysemy and salient reliance on context. [1]

A Mongolian NER model based on BERT-base tagging model. Eighteen types of fine-grained named entities related to tourism can be identified in the Mongolian Tourism Named Entity Recognition Corpus (MTNER).

Overall, our NER model extracted 18 travel-related named entity types, which scored 82.09% F1 in the Mongolian Tourism Named Entity Recognition Corpus (MTNER). This performance, we believe, is sufficiently strong to be practically useful. And we will release our data and code, including our annotated corpus, annotation guideline, a specially designed tokenizer, and a pre-trained Mongolian BERT and a trained NER model.

2 Related Work

Machine translation has been becoming more and more popular, especially in many special fields, including tourism. At the same time in the field of artificial intelligence knowledge also has great application prospects. [7]

The CoNLL 2003 dataset is a widely used bench-mark for the NER task. State-of-the-art approaches on this dataset use a bidirectional LSTM with the conditional random field and contextualized word representations.

Among all the methods, relay on the features to classify the input word, and do not count on linguists to make rules, supervised learning approaches have been the predominant in this filed. In the learning machine, each input will output a label with the learned algorithm, such as Hidden Markov Model [9], Support Vector Machine [10], Conditional Random Fields [11], and so on. Transfer-learning is also used for the NER task [2, 3, 26]. Various methods have been investigated for handling rare entities, for example incorporating external context or approaches that make use of distant supervision.

Named Entity Recognition has been explored for new domains and languages, such as social media, biomedical texts, multilingual texts, and the tourism domain. As to the techniques applied in NER, there are mainly the following streams. At first, rule-based approaches was the mainstream, which do not need annotated data as they rely on hand-crafted rules. Later, unsupervised learning approaches prevailed, which rely on un-supervised algorithms without hand-labeled training examples. Because of feature plays an vital role in the named entity recognition task, and feature-based methods become an inevitable trend, which rely on supervised learning algorithms with careful feature engineering. In recent years, with the development of deep learning, the method based on deep learning has become the mainstream, which automatically discover representations needed for the classification and detection from raw input in an end-to-end manner. [25]

In-domain research, the formerly Named Entity Recognition relay on feature engineerings, such as CRF and CRF-CRF [7,16] be used in the tourism domain, and much Statistical learning also uses into the tourism Named Entity Recognition, including HMM [17]. he latest research, the BERT-BiLSTM-CRF [7] be used in the Chinese military domain and got an excellent result. o we trained a NER model based on BERT-base tagging model for Mongolian in the tourism domain. n linguistics, a corpus or text corpus is a large and structured set of texts, and nowadays usually electronically stored and processed. What's more, in corpus linguistics, corpus is used to do statistical analysis and hypothesis testing, checking occurrences, or validating

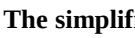
linguistic rules within a specific language territory [20, 24]. Consequently, it is crucial to build a quantity and quality corpus.

There has been relatively little prior work on named entity recognition in the tourism domain, use BERT-BiLSTM-CRF in Chinese tourism named entity recognition [19]. In this paper, we collected vast Mongolian data to pre-train a pre-training language model for Mongolian, built a corpus for Mongolian Tourism corpus, and annotated 18 types of travel-related entities to train a NER model base on the Mongolian Tourism corpus.

3 Challenge for Mongolian Tourism NER

In this section, we discuss the challenge for Mongolian language understanding and named entity recognition in tourism domain.

The named entities in the general domain, mainly including the names of the person, location, and organization name, have the characteristics of relatively stable type, standardized structure, and unified naming rules. While, in the tourism domain, named entities not only have the general domain challenges, including Large scale vocabulary, lack of abundant corpus, absence of capital letters in the orthography, multi-category word, subject-object-verb word order but also face other in-domain challenges, such as the entity boundary is not clear, the simplification expression, the rich entity types, the large quantity and so on.

The simplification expression. For example, "  " (Inner Mongolia University) also is said to "  ". Those phenomena increase the difficulty of identification.

The rich entity types. In the tourism domain, have many category entities, such as the display name in the scenic spot, is also the named entity should be annotated. It leads to many travel-related named entity need to recognize.

The large quantity. Various types named entity in the tourism domain, make the data quantity is large.

The research of NER in Mongolian started late and there are few related works, which largely restricted the development of informatization and intelligentization of Mongolian. In these years, NER has been emerging in the research on Mongolian language information processing. Significant achievements have been made in the identification of three categories of entities, person, location, and organization name. However, few actual achievements could have been made in the research on NER in other specific fields, including the tourism field.

In this paper, we collected a lot of tourism Mongolian data, and build an in-domain Mongolian tourism corpus for named entity recognition, meanwhile, we trained a Mongolian NER model based on BERT-base tagging model.

4 Annotated Mongolian Tourism Corpus

In this section, we describe the construction of our annotated Mongolian tourism NER corpus (MTNER). We collected Mongolian text data, and manually annotated them with 18 types of travel-related entities.

4.1 Data Collection

We collected Mongolian datasets, large and various, about 10GB, such as Mongolian news and Mongolian tourism, from many websites, including the Mongolian News website of China (http://www.nmgnews.com.cn/), Holovv (https://holoov.com), Ctrip (https://vacations.ctrip.com) and so on. Original datasets genres and sentences number in Table 1.

Table 1. Original datasets genres and sentences number.

Data genres	Sentence Number
News	24,593
Essays	256,959
Scenic Spot Intros	2,887
Travel notes	2,657
Others	1,744,681
Total	2,031,777

4.2 Annotation Schema

Based on the investigation and analysis, combined with the characteristic of the tourism domain, we find the traditional three categories, person name, location name, and organization name, is not enough, such as the location of tourism including the general location and scenic spot [7, 16, 17, 21].

Above all, we defined and annotated 18 types of fine-grained entities, including 2 types of Person entities, 6 types of Scenic entities, 4 tapes of Cultural entities, 4 types of Organization entities, and 2 types of Specific Field entities. The Person entities include mentions of Mongolian and Foreigner. The Scenic entities include mentions of Administration place, Natural sight, Public building, Marker building, Business, and Religion place. The Cultural entities include mentions of Culture, Education, Sports, and Musical production. The Organization entities include mentions of Company, Politics, and Charity. What's more, the Specific Field entities include mentions of Military, and Car, in Table 2.

We adopt BIOES Label schema, "B" represents the starting position named entity, "I" is inner a named entity, "E" means the ending position named entity, the single entity will be labeled "S". While, others will be labeled "O". That is all, we annotated 73 types of labels.

Table 2. Annotated entity classes and examples.

Coarse-grained	Fine-grained	Example	Means
Person	Mongolian	ᠬᠥᠳᠥᠯᠡᠭᠡ	Gentle
	Foreigner	ᠵᠡᠺᠺᠢ ᠨᠢᠺᠤ	Jack
Scenic	Administration place	ᠬᠥᠬᠡᠬᠣᠲᠠ	Hohhot
	Natural sight	ᠠᠶᠤᠯ ᠤᠨ ᠮᠣᠩᠭᠣᠯ	the Big Qing Mountain
	Public place	ᠪᠠᠶᠢᠲᠠ ᠣᠯᠠᠨ ᠤᠯᠤᠰ ᠤᠨ ᠭᠠᠽᠠᠷ	Baita International Airport
	Marker building	ᠬᠤᠪᠢᠯᠭᠠᠨ ᠪᠠᠷᠢᠭ	the Drum Tower
	Business	ᠲᠠ/ ᠤ/ ᠪᠠᠭᠠᠽᠠᠷ ᠤᠨᠠᠭᠠᠨ	the Wanda Squre
Cultural	Religion	ᠨᠤᠮ ᠰᠤᠭ ᠲᠡᠮᠳᠡᠭ	Jokhang Temple
	Culture	ᠠᠷᠠᠳ ᠤᠨ ᠰᠢᠨᠵᠢ	Tianbian
	Education	ᠮᠣᠩᠭᠣᠯ ᠤᠨ ᠤᠯᠠᠮ ᠵᠢᠩᠰᠢᠯ ᠦᠨᠡ	the Mongolian Traditional Culture
	Sports	ᠪᠦᠭᠡᠴᠡᠭᠡ	Wrestling
	Music	ᠶᠢᠭᠡᠳᠡ	Flute
Organizations	Department	ᠦᠪᠡᠷ ᠮᠣᠩᠭᠣᠯ ᠤᠨ ᠨᠤᠮ ᠦᠨᠳᠦᠰᠦᠯᠡᠭᠦᠢ	Inner Mongolia University
	Company	ᠬᠤᠸᠠ ᠸᠡᠢ ᠺᠣᠮᠫᠠᠨᠢ	HUAWEI
	Politics	ᠺᠣᠮᠫᠤᠳᠧᠷ ᠤᠨ ᠶᠡᠬᠡ	ACM
	Charity	ᠤᠯᠠᠭᠠᠨ ᠭᠠᠭᠠᠰᠠᠯᠠᠭ ᠤ ᠨᠡᠢᠭᠡᠳᠡᠯ	the Red Cross Society
Other Fields	Military	ᠲᠠᠩᠺ	Tank
	Car	ᠪᠡᠨᠽ	Mercedes-Benz

4.3 Annotation Agreement

Our corpus was annotated by eight annotators, who are college students, majored in computer science, are Mongols.

We used a web-based annotation tool, BRAT, and provided annotators with links to the original travel-related datasets of our collections. We adopted the cross-annotating strategy, four steps following:

1. We divide the data into eight parts and divided annotators into four groups.
2. Everyone annotated one part data.
3. Members of the same group should exchange data to cross-annotate. The inter-annotator agreement is 0.85, measured by span-level Cohen's Kappa (Cohen, 1960).
4. Manually check the marked results.

After those annotated operates, we got the annotated results, those can be saved to the .ann files, including four columns, ID, entity type, start position and end position, entity, in Fg.2.

T1	Education 3491 3509	《 ᠲᠤᠭᠠᠯ ᠤᠨ 》 ᠨ	(《Water bank》)
T2	Administration 2159 2176	ᠢᠲᠠᠯᠢ ᠤᠯᠤᠰ	(Italy)
T3	Foreigner 2388 2396	ᠰᠢᠯᠤᠬᠠ	(Mr. Shiloh)
T4	Foreigner 2572 2580	ᠰᠢᠯᠤᠬᠠ	(Mr. Shiloh)
T5	Foreigner 2696 2704	ᠰᠢᠯᠤᠬᠠ	(Mr. Shiloh)
T6	Mongolian 2762 2772	ᠽ · ᠴᠢᠯᠵᠢ	(Watts, Jill)
T7	Mongolian 3472 3490	ᠨᠠ · ᠰᠣᠩᠲᠤ ᠮᠤ	(Nathan Sontiian)
T8	Mongolian 3716 3734	ᠨᠠ · ᠰᠣᠩᠲᠤ ᠮᠤ	(Nathan Sontiian)
T9	Education 3735 3748	《 ᠲᠤᠭᠠᠯ ᠤᠨ 》	(《Water bank》)
T10	Mongolian 4624 4640	ᠠᠮᠢᠷᠠᠯᠲᠤ ᠳᠤᠭ	(Dag admiral)

Fig. 2. Annotated results, including ID, entity type, start position, end position, and entity. And the right-most column which we give the English translation of Mongolian.

5 Mongolian Tourism NER model

In this section, we introduce our Mongolian Tourism NER model with pre-training and fine-tuning strategy. We pre-trained BERT model for Mongolian, and fine-tuned a NER model based on BERT-base tagging model.

BERT, Bidirectional Encoder Representations from Transformers, is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks. One important aspect of BERT is that it can be adapted to many types of NLP tasks very easily [5].

Pre-trained Language model. We use collected Mongolian corpus, unlabeled, about 10GB, releasing the BERT-base, pre-trained a BERT model for Mongolian (Mongolian_base (12-layer, 768-hidden, 12-heads)) [12]. Training parameters in Table 4.

Mongolian Tourism NER model. The corpus was converted into BIOES label schema for fine-tuning the BERT-base model [12]. We trained our classifier task for NER base on the Mongolian Tourism corpus[13]. Training parameters in Table 3. Model structure in Fig.3.

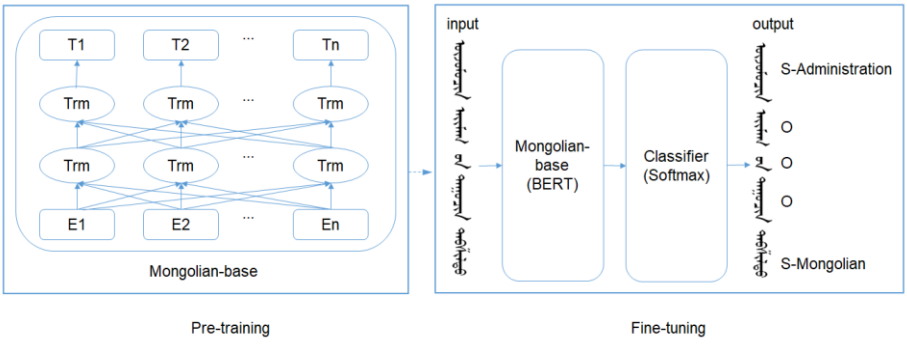


Fig. 3. NER model fine-tuned the BERT-base tagging model. Trained our NER model base on our Mongolian Tourism Named Entity Recognition corpus. Input a sentence, and output including each word label. Such as, in the sentence "ᠪᠤᠵᠤᠤᠷᠵᠤ ᠮᠤᠵᠢᠨ ᠳᠤ ᠠᠮᠢᠷᠠᠯᠲᠤ ᠠᠭᠤᠢᠯᠠᠭᠤ" , the word" ᠪᠤᠵᠤᠤᠷᠵᠤ " means "WuZhu MuQin", is a Location name and it only has one word which we annotated "S-

Administration" tag, and the word "ᠳᠠᠪᠤᠰᠢᠯᠠᠲᠤ" means "Dabuxilatu", is a Mongolian name and it only has one word which we annotated the "S-Mongolian" tag.

Table 3. Pre-training, Fine-tuning parameters, and values.

	Parameter	Value
Pre-training	max_sequence_length	128
	max_predictions_per_seq	20
	masked_lm_prob	0.15
	train_batch_size	8
	num_train_steps	200,000
	learning_rate	2e-5
Fine-tunning	train_batch_size	8
	eval_batch_size	8
	predict_batch_size	8
	learning_rate	5e-5
	num_train_epochs	3
	max_seq_length	128

6 Experiment

In this section, we show that our Mongolian NER model outperformance. To evaluate the Mongolian tourism domain named entity recognition model proposed in this paper, fine-tuned the BERT-base tagging model base on Mongolian Tourism Corpus named entity recognition (MTNER), and compared with those mainstream models for named entity recognition work with our corpus, including CRF and BiLSTM-CRF.

6.1 Data

The original Mongolian text data which we collected, have various problems, such as misspelling problem, we use the spelling correction to solve those errors [15], got the unannotated Mongolian corpus and the annotated Mongolian Tourism corpus for NER.

Mongolian corpus. We pre-trained a BERT model for Mongolian base on our Mongolian corpus, unannotated, about 10GB. We divided the data into three parts, train, development, and test set.

Mongolian Tourism corpus for NER. We train and evaluate our NER model on the Mongolian Tourism corpus of 12,800 train, 1,600 development, and 1,600 test sentences. We used the manually annotated corpus in (§4), it is a manually annotated Mongolian Tourism corpus, it contains 16,000 sentences and 15,320 named entities. The person, scenic, cultural, organization, and other fields named entities account for 22.56%, 32.53%, 15.25%, 20.36%, and 9.30%. The account of the fine-grained class in Table 4.

Table 4. The account of fine-grained class.

Entity type	Proportion (%)	Entity type	Proportion (%)
Mongolian	14.30	Education	4.62
Foreigner	8.26	Sports	2.37
Administration place	4.78	Music	2.63
Natural sight	5.74	Department	8.40
Public place	4.17	Company	3.26
Marker building	5.28	Politics	4.23
Business	4.25	Charity	4.47
Religion	8.30	Military	3.60
Culture	5.63	Car	5.70

6.2 Baselines

We compared our NER model with two mainstream named entity recognition models, our NER model outperformed than them.

A Feature-based Linear CRF model. This model uses standard orthographic, contextual, and gazetteer features. We implemented a CRF baseline model to extract the travel-related entities with the word-level input embedding. The regular expressions are developed to recognize specific categories of travel-related entities. [11] We use the "LBFGS" algorithm, and the cost parameters is 0.1.

A BiLSTM-CRF model. This model uses a BiLSTM-CRF network to predict the entity type of each word from its weighted representations. Using the contextual word embedding (ELMo) embeddings, and is used as the state-of-the-art baseline named entity recognition models in various special domains. [12] We set the word embedding size is 128, and the dimensionality of LSTM hidden states is 128. And set the same initial learning rate, batch size and epochs.

6.3 Results

We evaluated the results by the CoNLL metrics of precision, recall, and F_1.[6] Precision is the percentage of corrected named entities, recall is the percentage of named entities existing in the corpus and F_1 is the harmonic mean of precision and recall, these can be expressed as:

$$precision = \frac{Num(correct\ NEs\ predicted)}{Num(NEs\ predicted)}$$

$$recall = \frac{Num(correct\ NEs\ predicted)}{(\ Num\ all\ NEs\ )}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

On the same training set and test set, we compared the above two models. Table 5 shows the precision (**P**), recall (**R**), and $\mathbf{F_1}$ score comparison of the different models on our Mongolian Tourism corpus named entity recognition.

Table 5. Evaluation of the test sets of the Mongolian Tourism NER corpus.

	P (%)	R (%)	F_1 (%)
CRF	75.10	82.33	78.55
BiLSTM-CRF	76.32	84.25	80.08
Mongolian Tourism NER model	**78.59**	**85.94**	**82.09**

The results indicate that our Mongolian Tourism NER model is better than the other two named entity recognition models. Compared with the CRF named entity recognition model, BiLSTM can learn more contextual features. The model proposed in this paper improves the F_1 by 3.54% and the recall by 3.61%. Compared with BiLSTM-CRF named entity recognition model, our model improves the F_1 by 2.01% and the recall by 1.69%.

The features of word-level ignored the feature with the contextual, this model is a combination of words, sentences, and location features generated word representation, and using the Transformer to train the model, fully considering the influence of the contextual information of the entity, and got a better result.

6.4 Analysis

Pre-trained language model. Pre-training on large text corpora can learn common language representations and help complete subsequent tasks, s pre-training is an essential task in NLP. Pre-trained representations can also either be context-free or contextual, and contextual representations can further be unidirectional or bidirectional. Context-free models such as Word2vec, generate a single "word embedding" representation for each word in the vocabulary. Contextual models instead generate a representation of each word that is based on the other words in the sentence, such as ELMO, but crucially these models are all unidirectional or shallowly bidirectional. This means that each word is only contextualized using the words to its left (or right). Some previous work does combine the representations from separate left-context and right-context models, but only in a "shallow" manner. BERT represents one word using both its left and the right context, starting from the very bottom of a deep neural network, so it is deeply bidirectional. BERT outperforms previous methods because it is the first unsupervised, deeply bidirectional system for pre-training NLP. [5]

Training data scale. Usually, trains a pre-trained language model needs a large corpus. The corpus size of model training directly affects the performance of the model. The large scale of training data enables the model to fully learn the characteristics of language, to make full use of the corpus information to solve the problem of language understanding [20, 21]. So our data scale is not enough, we need to annotate more tourism and another domain corpus to support the downstream NLP task.

The proportion of entity categories. The text classification task, category distribution balance is very important to the classification model. Unbalanced classification makes it easy for the model to forget the categories that appear less frequently [2, 3, 18, 19, 21]. In our Mongolian Tourism corpus, the proportion of annotated entity types is balanced, it could trained a classifier model to be better.

7 Conclusion

In this work, we investigated the task of named entity recognition in the Mongolian Tourism domain. We collected a vast Mongolian text, developed a Mongolian Tourism Corpus of 16,000 sentences from the Mongolian Tourism domain annotated with 18 fine-grained named entities. This new corpus is an benchmark dataset for contextual word representations. We also pre-trained a BERT model for Mongolian and fine-tuned a NER model based on BERT-base tagging model for Mongolian Tourism named entity recognition. This NER model outperforms other mainstream NER models on this dataset. Our pre-trained Mongolian-base consistently helps to improve the Mongolian NER performance. We believe our corpus, BERT embedding for Mongolian, fine-tuned BERT-base tagging model for Mongolian Tourism NER model will be useful for various Tourism tasks and other Mongolian NLP tasks, such as Tourism Knowledge Graph, Mongolian Machine Translation, Mongolian question-answering, and so on.

8 Acknowledge

The project are supported by National Natural Science Foundation of China (No. 61773224); Inner Mongolia Science and Technology Plan(Nos. 2018YFE0122900, CGZH2018125, 2019GG372, 2020GG0046); Natural Science Foundation of Inner Mongolia (Nos. 2018MS06006, 2020BS06001); Research Fund for Inner Mongolian Colleges(No. NJZY20008); Research Fund for Inner Mongolian Returned Oversea Students and Inner Mongolia University Outstanding Young Talents Fund.

References

1. Tabassum J , Maddela M , Xu W , et al. Code and Named Entity Recognition in StackOverflow[J]. arXiv, (2020).
2. Wang, W, Bao, F. and Gao, G. Learning Morpheme Representation for Mongolian Named Entity Recognition. Neural Process Lett 50, 2647–2664. (2019).
3. Wang, W, Bao, F. and Gao, G. Mongolian Named Entity Recognition with Bidirectional Recurrent Neural Networks. The 28th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2016), pp 495-500. (2016).
4. Marcus MP, Marcinkiewicz MA, Santorini B, et al. Building a large annotated corpus of English: the Penn treebank[J]. Computational Linguistics. 19(2): 313-330. (1993).

5. Devlin J, Chang M, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. North American chapter of the Association for Computational Linguistics, 4171-4186. (2019).
6. Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvae Investigationes. 30(1): págs. 3-26. (2007).
7. Geng X. Research and Construction of the Map of Mongolian and Chinese Bilingual Knowledge for Tourism [D]. (2019).
8. Cao Y, Hu Z, Chua T, et al. Low-Resource Name Tagging Learned with Weakly Labeled Data[C]. international joint conference on natural language processing, 261-270. (2019).
9. Zhou G . Named entity recognition using an HMM-based chunk tagger[C]. North American chapter of the Association for Computational Linguistics '02 :Proc. 473-480. (2002).
10. Kudo T, Matsumoto Y. Chunking with support vector machines. North American chapter of the Association for Computational Linguistics.1508.01991. (2001) .
11. Lafferty, John and Mccallum, Andrew and Pereira, Fernando. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. 18th International Conf. on Machine Learning (ICML).282-289. (2002).
12. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer science, (2015).
13. Sun C, Qiu X, Xu Y, et al. How to Fine-Tune BERT for Text Classification?[J]. Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science,pp:194-206. (2019).
14. Yin X, Zhao H, Zhao J, Yao W, Huang Z. Named entity recognition in military field by multi-neural network collaboration [J]. Journal of Tsinghua university. 60(08):648-655. (2020).
15. Lu M, Bao F, Gao G, et al. An Automatic Spelling Correction Method for Classical Mongolian[J]. In: Douligeris C., Karagiannis D., Apostolou D. (eds) Knowledge Science, Engineering and Management. KSEM 2019. Lecture Notes in Computer Science, vol 11776. Springer, Cham. 201-214. (2019).
16. Guo J, Xue Z, Yu Z, et al. Named entity identification in tourism based on cascading Conditions [J]. Chinese Journal of Information Technology. 023(005):47-52. (2009).
17. Xue Z, Guo J, Yu Z, et al. Identification of Chinese tourist attractions based on HMM [J]. Journal of Kunming University of Science and Technolog. 34(006):44-48.(2009).
18. Li Dongdong. Named entity recognition for medical field [D]. (2018).
19. Zhao P, Sun L, Wan Y, Ge N. BERT+BiLSTM+CRF based named entity recognition of scenic spots in Chinese [J]. Computer system application. 29(06):169-174. (2020).
20. Wang C. The Research and Construction of Yi Corpus for Information Processing. 3(4). (2019).
21. Beibei Lin, Po-ching Yip. On the Construction and Application of a Platform-Based Corpus in Tourism Translation Teaching. 2(2):30-41. (2020).
22. Ren Z, Hou H, Jia T, Wu Z, Bai T, Lei Y. Application of particle size segmentation in the translation of Mongolian and Chinese neural machines [J]. Chinese journal of information technology. 33(01):85-92. (2019).
23. Cui J, Zheng D, Wang D, Li T. Entity Recognition for chrysanthemum named poems based on deep learning model [J/OL]. Information Theory and Practice. 1-11. (2020).
24. Liu G. Construction of parallel corpus for Legal Translation [J]. Overseas English. (10):32-33+39. (2020).
25. Li J, Sun A, Han J, et al. A Survey on Deep Learning for Named Entity Recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 1-1. (2020).

26. Wang W , Bao F , Gao G . Mongolian Named Entity Recognition System with Rich Features. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 505-512. (2016).