A BERT-Based Neural System for Chinese Short Text Entity Linking*

Chao Huo[1]✉, Xuanwei Nian[2], Deyi Xiong[3], Hanchu Zhang[1], Chao Wang[1], Changjian Hu[1], and Feiyu Xu[1]

[1] Lenovo Research, No. 10, East Xibeiwang Rd., Haidian District, Beijing, China
{huochao2,zhanghc9,wangchao31,hucj1,fxy}@lenovo.com
[2] School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
nianxuanwei218@buaa.edu.cn
[3] College of Intelligence and Computing, Tianjin University, Tianjin, China
dyxiong@tju.edu.cn

Abstract. In this paper, we present a Chinese short text entity linking system based on BERT (Bidirectional Encoder Representations from Transformers) [1]. BERT as a state-of-the-art pre-training language model has achieved promising results in many language understanding tasks. In particular, our task is mainly composed of two parts, entity recognition and entity disambiguation. For each part, we use BERT re- gression model for fine-tuning. Our approach achieves F1-score of 79.96% on the final test data which is ranked in the second place in the contest of CCKS 2019 ERL (Entity Recognition and Linking) task.

Keywords: Entity Linking · BERT · transfer learning · short text.

1 Introduction

Entity Recognition and Linking (ERL) is one of the basic tasks in the Natural Language Processing (NLP) field, which is, for a given Chinese short text (such as search Query, Weibo, user dialogue content, article title, etc.) identifies the entities in it and associates them with the corresponding entities in a given knowledge base. The entire process of ERL includes two subtasks: entity recognition and entity linking two subtasks.

The traditional entity linking refers to the task mainly for long documents, and the long document has the rich context information to assist the entity's disambiguation. In contrast, the entity linking of Chinese short texts has great challenges. The main reasons are as follows: (1) Serious colloquialism, which makes the disambiguation of entity ambiguity difficult; (2) The context in short text is not rich, and we must understand the true meaning of context accurately; (3) Compared with English, Chinese is more challenging in the short-text entity linking task because of the language's own characteristics.

* Supported by Lenovo AI-Lab.

A key part of applying deep learning methods to tackle this problem is the text representation. One kind of pre-trained models is the word embeddings, such as word2vec [2] and GloVe [3], or the contextualized word embeddings, such as ELMO [4]. These word embeddings are often used as additional features for the main task. More recently, pre-trained language models have shown to be useful in learning common language representations by utilizing a large amount of unlabeled data: e.g., OpenAI GPT [5] and BERT [1]. BERT is based on a multi-layer bidirectional transformer [6] and trained on plain text for masked word prediction and next sentence prediction .

We present a neural approach to discovering gold mentions and linking them to the correct entities in the given knowledge base. Firstly, we use n-grams to find all possible candidates, and some of them will be kept if they are in the knowledge base. Then, we use BERT to encode the contexts where the candidate occurs. After that, each possible span will get a score estimated by BERT, and we use the validation data set to find the optimal threshold. Each candidate which has a score greater than the threshold will be treated as gold mention and delivered to the next step. In the entity disambiguation part, our methods take advantage of all aspects of an entity, such as contexts around it, encyclopedia entity descriptions and entity type information.

2 Related Work

2.1 Entity Linking

Entity Linking (EL) is the task of recognizing (cf. Named Entity Recognition) and disambiguating (Named Entity Disambiguation) named entities to a knowledge base (e.g. Wikidata, DBpedia, or YAGO). It is sometimes also simply known as Named Entity Recognition and Disambiguation. In general, there are two main approachs for EL: End-to-End methods take plain text as input, extract gold mention at first and then link them to the correct entity in the given knowledge base; Disambiguation-Only methods, contrary to the first approach, directly take gold standard named entities as input and only disambiguates them to the correct entity in a given knowledge base.

2.2 Pre-trained Language Model

Pre-training on a large amount of unlabeled data and fine tuning in downstream tasks has achieved significant improvements on several natural language understanding tasks. BERT is trained on a large amount of cross domain corpus for next sentence prediction and masked language model task. Unlike previous bidirectional language models (biLM) limited to a combination of two unidirectional language models (i.e., left-to-right and right-to-left), BERT uses a Masked Language Model to predict words which are randomly masked or replaced. Experimental results have demonstrated impressive gains in a broad range of NLP tasks, from sentence classification to sequence labeling.

The BERT modeling object focuses on the original language signal and uses less semantic information. This problem is more significant in Chinese language filed. Baidu proposed an ERNIE (Enhanced Representation from Knowledge Integration) [7] model based on knowledge enhancement. ERNIE learns the semantic representation of the complete concept of the text by masking the semantic units such as phrases and entities. Recently, an upgraded version of BERT has been released with Whole Word Masking (WWM), which mitigate the drawbacks of masking partial WordPiece tokensin pre-training BERT, Yiming et al proposed the Chinese BERT-WWM [8] version.

3 BERT for Text Classification

BERT-Base-Chinese model contains 12 transformer blocks, 12 self-attention heads and hidden size is 768. The input sentence's size is restrict to no more than 512, and the output is a 768 dimensional vector. BERT has two special tags, [CLS] contains the special classification embedding and another special token [SEP] is used for separating segments.

Specific to the entity recognition and linking task, we use the regression model rather than classification model due to unbalanced training data. We take the final hidden state h of the first token [CLS] as the representation of the whole sequence. And we use mean square error as loss function.

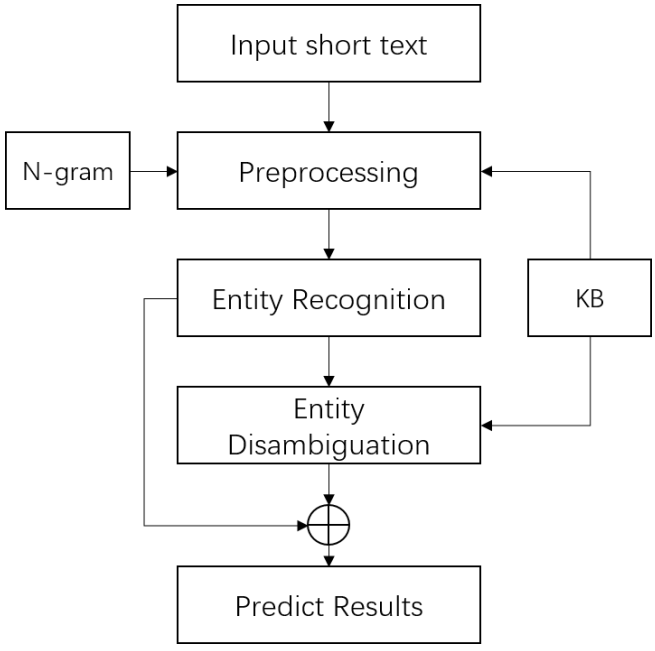

Fig. 1. Architecture of our System.

4 Methodology

The architecture of our system is shown in Fig. 1. We preprocess the input short text to get n-gram ($n = 1, 2, 3, ...$) words filtered by the given knowledge base. We use BERT to determine whether a n-gram phrase is a gold mention or not. BERT takes the original short text and n-gram word as input, and output the probability as gold mention. In the entity disambiguation module, we semantically match the entity description from knowledge base with the short text and choose the one with highest score as the correct entity.

4.1 KB Dictionary and Word Segmentation

In this paper, we construct a word dictionary and an entity description dictionary with the knowledge base (KB) provided by the organizer. The word dictionary is composed of all mentions and the content in entity description dictionary is the fusion of all attribute information of entity. The entity candidate word set is obtained by using the n-gram word segmentation method and the entity dictionary for filtering.

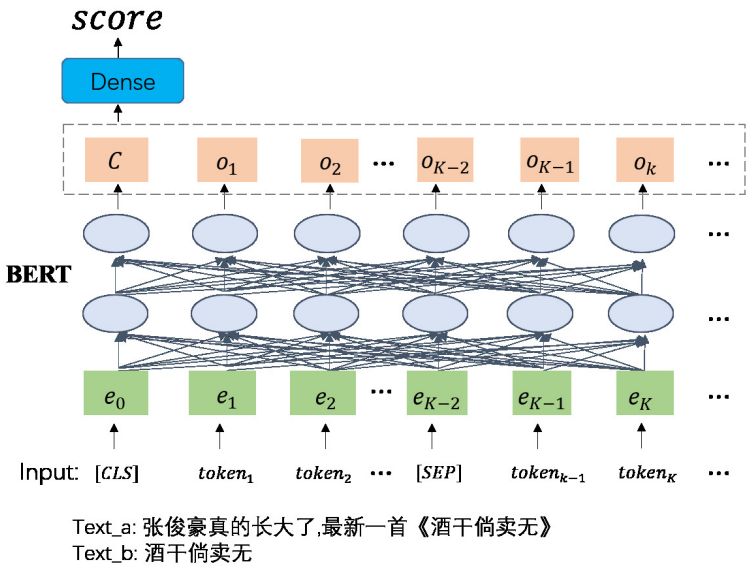

Fig. 2. The architecture of the entity recognition model.

4.2 Entity Recognition

Instead of applying the common sequence labeling model for entity recognition, we convert this task into a sentence-to-relationship judgment problem by fine-

tuning BERT pretrained model, which is more consistent with the BERT pre-training process. As illustrated in Fig. 2, at the input, Text A is the original short text and Text B is the candidate word. At the output, the high-level [CLS] representation is fed into an output layer for regression so that we can choose the best threshold based on the validation set for entity recognition.

4.3 Entity Disambiguation

In order to improve the accuracy of the entity linking, we use the all attribute information of the entity in the knowledge base and encode it through BERT to obtain the semantic space representation. The description is a combination of entity-specific attributes, including entity types, entity summaries, etc. Since the length of entity description maybe more than 512, so we put the entity type information in front of the entity description. The reason is the entity type information is more important for the entity disambiguation. And this trick makes the F1 score increased by 0.6%. As same as entity recognition task, we still use BERT regression model for the entity disambiguation, we take original text as Text A and entity description as Text B. Unlike entity recognition, at this point we don't need to find the optimal threshold with the validation set. We simply take the entity with the highest score as the result of the linking. The model architecture is shown in Fig. 3.

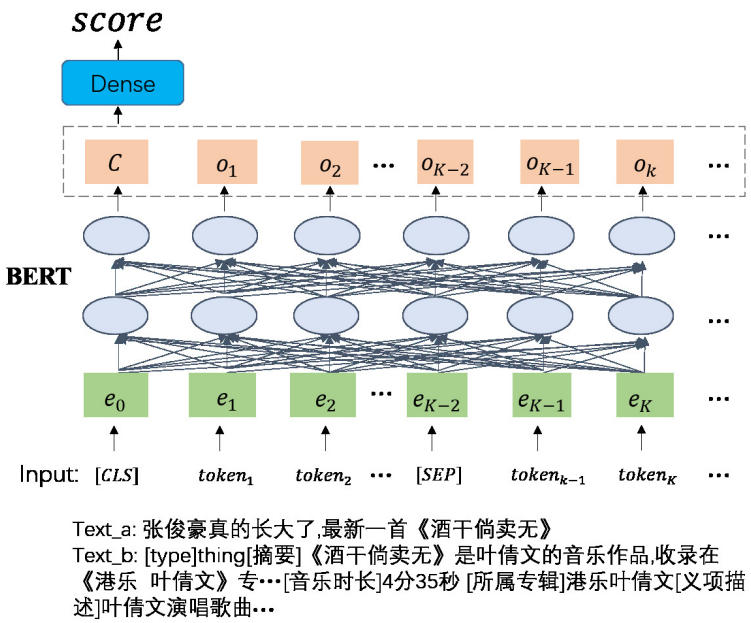

Fig. 3. The architecture of the entity disambiguation model.

5 Experiments

Our experiments are also composed of two parts, for the entity recognition part, we tried sequence labeling models and text classification models, but for the entity disambiguation part, we only tried BERT based model.

5.1 Datasets

We evaluate our method on the CCKS 2019 ERL task, the data set is composed of a knowledge base, training set,develop set and test set. The knowledge base includes approximately 390,000 entities from the Baidu Encyclopedia Knowledge Base. Each entity in the knowledge base contains a KB-ID, a string name, the type information, and a series of triplet ¡subject, predicate, object¿ information forms associated with the entity. The knowledge base data is distributed as follows:

Table 1. Statistics of the Knowledge Base

Type	Quantity
Number of entities	398,082
Number of SPO	3,564,565
Entity description data	361,778
Average number of entity attributes	9
Average length of entity description	103

5.2 Hyperparameters

We use the BERT-Base Chinese model and BERT-WWM and Ernie model,the last of two only need to download model weights and other settings are as same as BERT. We fine-tune the BERT model on tesla V100 GPU and with a batch size of 32, max sequence length of 512, learning rate of 2e-5, but for entity recognition task the max sequence length is set to 80. The dropout rate is set to 0.1, and we use Adam for optimization.

5.3 Entity Recognition Result

We tried two kinds of models, one is traditional sequence labeling model and the other is text classification model. For the sequence labeling model, we tried BiLSTM-CRF [9], BERT-BiLSTM-CRF, BERT-BASE,BERT-WWM-BASE and ERNIE-BASE. For the text classification model, we tried TextCNN [10], DPCNN [11], BERT-Regression, BERT-Regression-Ensemble. The experiment results are shown as follows.

Table 2. Entity recognition result in develop set

Tasks	Model	F1
Sequence Labeling	Bi-LSTM-CRF	75.65%
	BERT-Bi-LSTM-CRF	77.82%
	BERT-BASE	79.17%
	ERNIE-BASE	79.91%
Text Classification	TextCNN	72.03%
	DPCNN	75.88%
	BERT-Regression	84.51%
	BERT-Regression-Ensemble	**86.24%**

5.4 Entity Linking Result

The final entity linking result is evaluated on the biendata platform, our n-gram word segmentation model achieved 77.79% on the develop set and 79.97% on the test set.

Table 3. Final Result on biendata

Dataset	F1
Develop Set	77.79%
Test Set	79.97%

6 Conclusion

In this paper, we have proposed a BERT fine-tune method on the Chinese short text entity recognition and linking task. We achieved 79.97% F1-score on the final test set on CCKS 2019 ERL task. In the future, we will try the joint training methods instead of our current two step method.

Acknowledgements

Deyi Xiong was supported by National Natural Science Foundation of China (Grant Nos. 61622209 and 61861130364)

References

1. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

2. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13), C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 2. Curran Associates Inc., USA, 3111-3119.
3. Pennington, Jeffrey & Socher, Richard & Manning, Christoper. (2014). Glove: Global Vectors for Word Representation. EMNLP. 14. 1532-1543. 10.3115/v1/D14-1162.
4. Peters, Matthew & Neumann, Mark & Iyyer, Mohit & Gardner, Matt & Clark, Christopher & Lee, Kenton & Zettlemoyer, Luke. (2018). Deep Contextualized Word Representations. 2227-2237. 10.18653/v1/N18-1202.
5. Radford, Alec. Improving Language Understanding by Generative Pre-Training. (2018).
6. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 59986008.
7. Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., & Wu, H. (2019). ERNIE: Enhanced Representation through Knowledge Integration. ArXiv, abs/1904.09223.
8. Cui, Yiming & Che, Wanxiang & Liu, Ting & Qin, Bing & Yang, Ziqing & Wang, Shijin & Hu, Guoping. (2019). Pre-Training with Whole Word Masking for Chinese BERT.
9. Huang, Z., Xu, W.L., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. ArXiv, abs/1508.01991.
10. Kim, Yoon. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 10.3115/v1/D14-1181.
11. Johnson, Rie & Zhang, Tong. (2017). Deep Pyramid Convolutional Neural Networks for Text Categorization. 562-570. 10.18653/v1/P17-1052.