EasyChair Preprint

№ 3394

The Exploration of the Reasoning Capability of BERT in Relation Extraction

Lili Li, Xin Xin and Ping Guo

May 13, 2020

The Exploration of the Reasoning Capability of BERT in Relation Extraction

1st Lili Li *
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
3120191016@bit.edu.cn

2nd Xin Xin *
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
xxin@bit.edu.cn

3rd Ping Guo
School of Systems Science
Beijing Normal University
Beijing, China
pguo@ieee.org

***Abstract*—Relation classification task is to predict the relation between the entity pair in a given sentence. Most of these sentences have certain words or schema that can help to extract the relationships of entity pairs. However, there are some sentences do not have such structure, they require the model to have certain reasoning capability to predict relation correctly, we call them "reasoning instances". BERT is a well known pre-trained language model, which can learn text representation and has already performed well on various tasks of NLP. In this paper, we are intended to explore the reasoning capability of BERT in reasoning instances. We first propose a BERT-based relation classification model based on the MG Lattice model to test whether BERT could infer the relation between entities in reasoning instances correctly. Further we explore what kind of information would help BERT to predict the relation of these instances. Through various comparison experiment, we conclude that BERT can not infer the relation between entities by the meaning of the sentence, it mainly uses the concept information about the entity itself and the information learned on previous instances to help the model to do relation classification. The conclusion inspires us that BERT can serve to predict the relation between entity pairs defined by multiple sentences.**

***Index Terms*—BERT, relation classification, BLSTM, relation reasoning**

I. Introduction

For a given sentence, relation classification mainly to predict semantic relations between entity pairs in this sentence. For example, in the sentence "美国首位女国务卿 The first female secretary of U.S.A. ", the goal of the relation classification model is to predict the relation between "国务卿 secretary"and "美国 U.S.A. "is "Employment". Relation classification is a key module for constructing large-scale knowledge graphs, and places an important role in information extraction (IE). Currently with the development of deep learning, the neural relation extractions (NRE), which means apply the neural network such as RNN [1],LSTM [2],CNN [3] on relation extraction, has become one of the most popular topics in natural language processing (NLP).

* Corresponding author.

ACE Corpus 2005 contains six general relations and 18 relation subtypes for the relation classification task. We found that most instances in ACE have certain schema or words that represent the relationship between entity pairs, which can help deep learning network to capture essential information for relation extraction task. But for the remaining about 10% instances in ACE don't have such structure, they require the classification model to understand the meaning of the entire sentence and use this meaning to infer the relationships between the entity pairs. We call these instances as "reasoning instances". More details about these reasoning instances are introduced in section "Methodology".

Besides, multiple sentence instances or document level instances may be required when extracting the relationship between entity pairs. Therefore, an excellent relation extraction model not only uses the information of the current sentence but also keeps the entity information that learned before. It puts forward higher requirements for the deep learning relation extraction model.

Pre-trained language model, like word2vec [4], Glove [5],BERT [6] can learn word representations from their contexts and have proven to be effective against many NLP tasks as well as relation extraction. Bidirectional Encoder Representation from Transformers (BERT) [6] as one of the most effective pre-trained language models at present, which can fully describe character-level, word-level, and sentence-level features of the text. It has achieved outstanding performance in most areas of NLP, as well as relation classification task [7] [8].

Despite the success of these well pre-trained language models, most of them only use simple contextual features of the text. Therefore, many researchers believe that these pre-trained language models could not get enough information to understand the meaning of sentences. Although BERT can enrich the word representation by fully using the information of the context, it still doesn't take into account the structure information of the whole sentence. Therefore, we consider it is difficult for BERT to understand the meaning of the text, which will restrict the reasoning capability of BERT on the relation extraction task.

In this paper, we mainly to explore what kind of information would help BERT to conduct the relation extraction task and whether this information can help BERT infer the relationships between entity pairs. We first proposed a BERT based relation classification model which builds on Li's MG lattice model [9]. We use BERT to enrich the character-level representation of the MG lattice model and the experiment results show that our BERT-based MG lattice model has a significant improvement over the original MG lattice model.

Then we construct five kinds of test instance datasets to evaluate whether our BERT-based relation classification model has the reasoning capability and what kind of information BERT mainly uses when extracting relation from reasoning instances. We found that there are almost 800 reasoning instances in ACE Corpus, so we choose 200 reasoning instances and 20% of the original instances on ACE Corpus to create five different kinds of test datasets, which can conduct various compare experiments to explore the reasoning capability of BERT.

- **Relation classification test dataset** evaluates the performance of our BERT-based model on the whole ACE Corpus.
- **Reasoning test dataset** shows the power of our BERT-based relation classification model for the reasoning instances.
- **Relation representation test dataset** explores whether the BERT-based classification model has a limitation in understanding the meaning of the whole sentence.
- **Entity concept test dataset** confirms that BERT mainly uses the entity conceptual information to enrich word representations.
- **Specific replacement test dataset** indicates BERT also use the information learned before to enrich the present word representation, which can help us do more improvement in the relation classification task.

More details about the test datasets can be found in "Methodology".

The main contributions of our paper are as follows:

- We proposed a BERT-based relation classification model based on Li's MG lattice model [9]. And our model achieves quite an improvement over the MG lattice model.
- We construct five test datasets, which consist of the original test dataset, the reasoning test dataset, the relation representation test dataset, the entity concept test dataset, and the specific replacement test dataset.
- By comparing with the experimental results of these five datasets together, we conclude that BERT can' t use the meaning of sentences to infer the relation between entity pairs. BERT mainly uses the concept information of entities and the information that learned in previous sentences to enrich the current word representation. Since BERT could keep the information that has appeared before, it inspires us BERT can be utilized to predict the relation between entity pairs that defined by multiple sentences.

II. Related work

A. Traditional methods of relation classification

There are five different kinds of traditional methods of the relationship classification task, supervised-based, semi-supervised, unsupervised, distantly supervised, and open domain-oriented extraction.

We focus on the supervised relation extractors, they treat the relation extraction task as the classification problem and focus on using the sufficiently large labeled dataset to train a suitable classifier. So relation extraction and relation classification mean the same task in this situation. The most prevalent methods of supervised relation extraction are the feature-based and kernel-based methods.

The feature-based methods aim to train the suitable classifier as the final relation classification models through extracting different kinds of features from sentences [10] [11] [12] [13]. These methods can predict the relation between entity pairs from the reasoning instances by extracting the structural features of sentences. Kambhatla [10] combined various vocabulary, syntactic and semantic features of the text and used the maximum entropy classifier to construct the model. By adding these various types of text features, the model can learn enough information to predict the relationship between entities without constructing the semantic feature extraction trees. Culotta [13] treated the relation extraction task as a sequence labeling task and used MALLET CRF with default regularization parameters to make use of structural information from sentences. Most of these feature-based methods mainly focus on selecting more suitable features from sentences with the help of traditional machine learning tools.

Kernel-based methods [14] mainly use kernel functions to measure the similarity between objects to complete the classification task. The functions of kernel-based methods can make full use of long-distance features in the sentences, hence they are capable of capturing the pair-wise relationship between entities in reasoning instances. Zelenko [15] combined 5 basic and 2 composite kernel functions to map the various text features from low-dimensional space to high-dimensional space, so the nonlinear relation extraction task can be treated as a linear problem. Zhang [16] proposed a parse tree-based convolution kernel function that could learn the syntactic structure information of sentences. However, kernel-based methods always have a high time complexity, they are too costly for processing big data.

B. Deep Learning Network methods

A number of studies based on deep learning models have been proposed for relation classification task, such as [1] [2] [3]. They applied various classic neural networks such as recurrent neural network(RNN), convolutional neural networks (CNN), bidirectional long short term memory (BLSTM) to automatically extract the context, semantic and structural features of sentences. Socher [1] applied RNN and parse trees to capture the compositional meaning of longer phrases. Zeng [17] used a convolutional deep neural network to automatically extract hierarchical features of sentences for relation extraction, which reduced the complex syntax and semantic processing. In addition to BLSTM, Zhou [2] also added the word level attention mechanism to improve the relation classification.

However, all these above relation classification methods require marking the position of the entity pair in the input text. The Named Entity Recognition(NER) task which aims to identify entities with specific meaning in the text is the pre-task of the relation classification task. Some researchers joint these two tasks together and train the NER model and RE model simultaneously to make use of the relationship between the NER task and RE task. The multi-task methods also reduce the risk of error accumulation in previous pipeline extraction. Ye [18] proposed an end-to-end joint extraction model to extract entities and their relationships at the same time, these two tasks share the parameters and optimization. In addition, they use the BIO tag of NER to enrich the input representation of relationship extraction and have made significant improvements in ACE 2005.

Due to the difficulty of Chinese word segmentation and the lack of corpora and NLP tools, there are relatively fewer papers discussing Chinese relation extraction compared with the English. Zhang [19] combined the shortest dependency path (SDP) with LSTM to extract semantic features from Chinese sentences to predict the relationship between entities. For the problem that most of the existing methods for Chinese relation classification do not consider the different granularity of the input may affect the model significantly, Li [9] proposed a multi-grained lattice framework (MG lattice) MG lattice model that fully captures both word and character-level features by introducing external linguistic knowledge HowNet, which can avoid the segmentation errors. It significantly outperforms multiple existing methods, achieving state-of-the-art results on ACE Corpus 2005. And our work is built on this MG lattice model.

C. Reasoning on relation classification task

For weakness that deep learning can't do causal reasoning, many researchers use graph neural networks to overcome it recently. Graph neural network(GCN) can extract the structure information of text to help deep learning networks perform better in various NLP tasks.

DeepMind proposed the relation network(RN) module in 2017 [20]. The neural network with the RN module has the capability to process unstructured input. For a given picture or sentence, the model can infer the relations among them. The model has achieved excellent results in visual QA and based QA.

Then in 2018, 27 authors from DeepMind, Google Brain, MIT, and other institutions proposed a "graph network" [21], which promoted and expanded GNN, making GNN have a strong relationship induction bias. Graph network promotes the extraction of structural information in the text and provides a new possibility for relational reasoning.

Currently, lots of researchers have used Graph Neural Network to do the NLP task, as well as the relation classification task.

Zeng [22] combined GCN with the dependency tree to do the relation extraction task. They first pruned the dependency tree by using a rule-based method, so that the only words in the shortest path between two possible entities were included in the tree. Then, these pruned trees were used as input of the GCN network to extract structured information from the input text. The model can use the dependency structure on the input statements to obtain the non-local syntactic relationship which is difficult to understand from the surface form. However, if the edge does not appear in the shortest path, its weight will be recorded as 0 when pruning the dependency tree, which means that sometimes important information in the sentence will be lost due to excessive pruning.

Guo [23] put forward a "soft pruning" strategy based on Zeng's model [22]. In order not to lose information in the sentence, it uses a fully connected edge-weighted graph as full dependency tree to replace the original dependency tree and utilize the self-attention mechanism to learn the strength of relatedness between nodes. Then they use these full dependency trees as the input of GCN to learn more expressive representation. Since the GCN can use the full dependency trees, it can extract more structure information from sentences to predict the relationships of the entity pairs.

All of these methods use graph neural networks to capture the structure information of sentences to infer the relation between entity pairs. As a pre-train language model, BERT can fully learn contextual features of sentences, so whether Bert can replace GCN to extract the structure information of sentences is an issue worth exploring.

D. BERT for relation extraction task

Bidirectional Encoder Representation from Transformers (BERT) [6] is a well-know pre-train language model that can fully describe character-level, word-level, and sentence-level features after learning contextual words from a large amount of training data. BERT has achieved outstanding performance in most areas of natural language processing. For the relation classification task, Wu [7] first

proposed the relation extraction model that using the pre-trained BERT language model and achieved improvement over the state-of-the-art method on the SemEval-2010 task 8. Soares [8] used different input and output strategies of BERT to explore how to make BERT perform better on relation classification.

Despite BERT as a well pre-trained language model performs excellently in the above models, we argue that it only focuses on language modeling and this may restrict the power of the pre-trained representations. Although BERT can enrich the word representation by fully using the information of the context, it can not learn the structure information of the whole sentence. It uses simple contextual features for word representation and training targets, which could not be enough to understand the meaning of sentences. Thus we assume that BERT could not infer the relationship between entity pairs by understanding the meaning of the sentence.

III. Methodology

We found there are about 10% instances in ACE Corpus 2005 require the RE models to have considerable reasoning capability to predict the relationships between entities, we call them "reasoning instances". To verify the reasoning ability of BERT, we propose a relation classification model based on BERT, whose f1 value exceeds the original MG lattice model by about 4.6%. We first use the reasoning instances to construct five different kinds of test datasets and then conduct various comparison experiments to explore whether our BERT-based relation extraction model could infer the relation between entities based on understanding the overall meaning of the sentence. Besides we continue to explore what kind of information that BERT uses to enrich the word representation in relation extraction task.

A. reasoning instances

We have found a rule that the relationship between many entities can be expressed by certain words or schema. For example, in the phrase "Minister of China's personnel department 中国人事部部长", the word "minister 部长" could illustrate that the relationship between "Minister of China's personnel department 中国人事部部长" and "China's personnel department 中国人事部" is "employment". These types of instances account for about 90% in the ACE Corpus 2005.

But for the remaining 10% of the instances, no word or schema can directly indicate the relationship between entities in the sentences. The relation extraction model is required to have a certain reasoning ability to correctly predict the relationship between entity pairs. For example, "DengFeng is the host city, because Shaolin Temple is the origin of Shaolin Boxing 登封市也因为少林拳的出处少林寺的缘故，成为主办城市", although no word or schema can indicate the relation between "登封 DengFeng "and "少林寺 Shaolin Temple ", we can infer the relationship between "DengFeng " and "Shaolin Temple" is "Location" from the following steps:

- First we should understand the meaning of this sentence is "DenFeng"could be the host city of the event because "Shaolin Temple"is located in "DengFeng".
- Then we can find that in the meaning of the sentence, the word "located" can indicate the relationship between these two entities is "location".

So that we can infer the relation between "DengFeng" and "Shaolin Temple"is "location"based on understanding the meaning of sentences. We call that reasoning capability. And the instances which require reasoning capability to do relation classification, we call them reasoning instances. The difference between reasoning instances and other instances is shown in Figure 1.

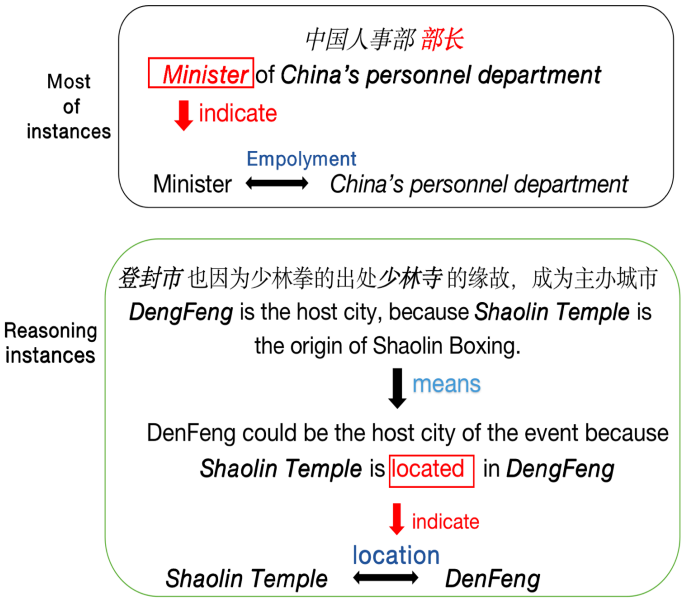

Fig. 1. The different processes for predicting entity relationships at most of the instances and reasoning instances on ACE.

Even if reasoning instances only account for 10% on the whole dataset, they become the bottleneck of the relation classification task. How to improve the deep learning network to correctly classify these instances is an important issue in NLP.

For the pre-train language model, like word2vec, Glove, and BERT, they mainly use the contextual information to obtain vector representation of words. The more context information they use, the better the model performance will be. So when using BERT as the pre-train language model to train our relation classification model, it will learn more information from the instances which have certain words or schema to indicate the relationship between the entities pairs. Because BERT can easily obtain entity information from the context of these instances. However, only use the contextual information of the word is difficult to capture the structure information of the

whole sentence. We assume that the pre-train language model BERT cannot fully understand the meaning of sentences, which means that BERT couldn't use the meaning of the sentence to infer the relations in the reasoning instances as well. To confirm our assumption, we use the reasoning instances mentioned above to construct various test datasets to explore whether BERT can learn the meaning of sentences, and what information BERT will use to perform the relation classification task.

B. Create five kinds of test data

As mentioned above, we found that there are almost 800 reasoning instances in ACE Corpus 2005 for the relation classification task. Based on this characteristic of the dataset, we create five different kinds of test datasets to explore the reasoning capability of BERT and discuss what kind of information does BERT use to enrich the word representation further.

First of all, we divided the 800 reasoning instances into two parts, including 600 reasoning instances as the training dataset and the other 200 reasoning instances as reasoning test datasets. The remaining 90% of the ACE Corpus 2005 instances are divided into the training dataset and the test dataset at a ratio of 8: 2. Then we combine the 600 reasoning training instances with the remaining training datasets of ACE Corpus mentioned above as the final training dataset to train our BERT-based model.

Next, the following five different kinds of test datasets will be constructed to explore the reasoning capability of BERT through the performance of our BERT-based relation classification model, which is the main work of our paper:

The first relation classification dataset. We use this dataset to evaluate the performance of our BERT-based model on the relation classification task. We use the 20% of the original ACE Corpus 2005 dataset mentioned above to be the first test dataset in our paper. Each of these instances has certain schema, words, or phrases to indicate the relation between entity pairs. For example, "President of Russia", the word "President"indicates that the relation between "President of Russia "and "Russia" is "Employment". We except that our BERT-based deep learning relation classification gets the highest F1-score on this test dataset than another four datasets.

The second reasoning test dataset. It consists of the 200 reasoning instances, including all 18 subtypes of ACE Corpus 2005. An example of this dataset is shown in Figure 2.

If the F1-score of our BERT-based model on this reasoning test dataset is worse than the first relation classification test dataset, we can conclude that the BERT-based MG lattice model has a limitation in understanding the meaning of the whole sentence. That is to say, BERT does not have the reasoning capability. However, since BERT can capture much information from character level, word level and sentence level from the sentences to enrich the word representation, although the model could not understand the entire meaning of the reasoning test dataset, we also assumed that the F1-score may not drop sharply.

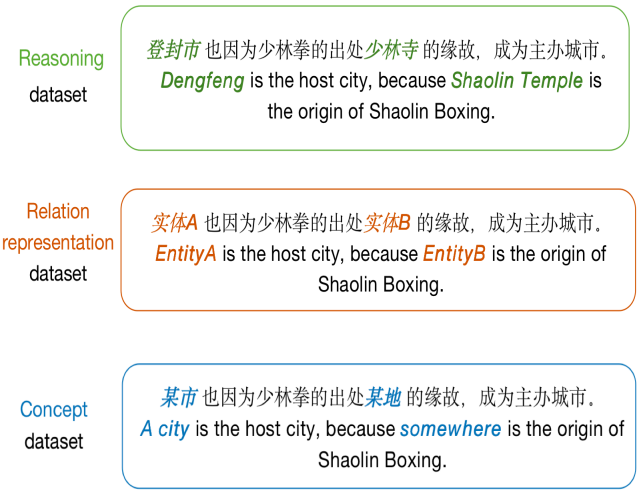

Fig. 2. Examples of reasoning dataset, relation representation dataset and concept dataset.

The third relation representation test dataset. We call it the "relation representation dataset"because this dataset is inspired by the Soares's [8] definition of the relation representation. In this test dataset, we replaced the entity pairs by "实体 A Entity A"and "实体 B Entity B"in the second reasoning test dataset. Compare with the second reasoning test dataset, we get 200 instances like "实体 A 也因为少林拳的出处实体 B 的缘故，成为主办城市 Entity A is the host city, because Entity B is the origin of Shaolin Boxing". The difference between these two datasets is also shown in Figure 2.

We mask the entity information in this test dataset to explore whether the BERT-based model could infer the entity relationships without learning any information from the entity pairs. Since the entity pairs are masked by "EntityA 实体 A" and "EntityB 实体 B", BERT could not use the information from entity pairs themselves to enrich the word representations in sentences. In this situation, if our BERT-based model could understand the whole meaning of the sentence, it can still infer the relationship between "EntityA 实体 A" and "EntityB 实体 B" by the sentence meaning.

In this paper, we assumed that BERT could not understand the meaning of sentences because it can not capture the structural information from sentences. So we estimate the F1-score of our BERT-based model on this relation representation test dataset will drop sharply. And the experiment on section 4 has confirmed our assumption.

The forth concept test dataset. Since we confirmed that BERT cannot infer the relation of the reasoning instances by the relation representation, how could our BERT-based model get a quite excellent result on the

second reasoning test dataset? To answer this question, we construct the fourth concept test dataset. The examples are shown in Fig.2. We know that the information contained in the entity itself will also affect the relationship. So we conduct the forth concept test dataset base on the third relation representation test dataset. We replace the mask entity pairs with the words which can indicate the concept information of the entity. For example, the original reasoning instance is "DengFeng is the host city, because Shaolin Temple is the origin of Shaolin Boxing 登封市也因为少林拳的出处少林寺的缘故，成为主办城市", and it is masked as "实体 A 也因为少林拳的出处实体 B 的缘故，成为主办城市 Entity A is the host city, because Entity B is the origin of Shaolin Boxing"in the third relation representation test dataset. Then in this concept test dataset, we use the concept words "A city" and "somewhere" to replace the "EntityA" and "EntityB" respectively, and the instance on this dataset has become " 某市也因为少林拳的出处某地的缘故，成为主办城市 A city is the host city, because somewhere is the origin of Shaolin Boxing".

"A city" and "somewhere" retain the concept information of the entities and we believe that BERT can use the concept information of entities to greatly enrich the word representation for the relation classification task.

If the F1-score of this test dataset improves over the third relation representation test dataset, we can conclude that although BERT can not infer the entity-relationship from the reasoning instances through the meaning of the sentences, it still obtains a quite excellent result on the second reasoning test dataset due to the use of the concept information of entities.

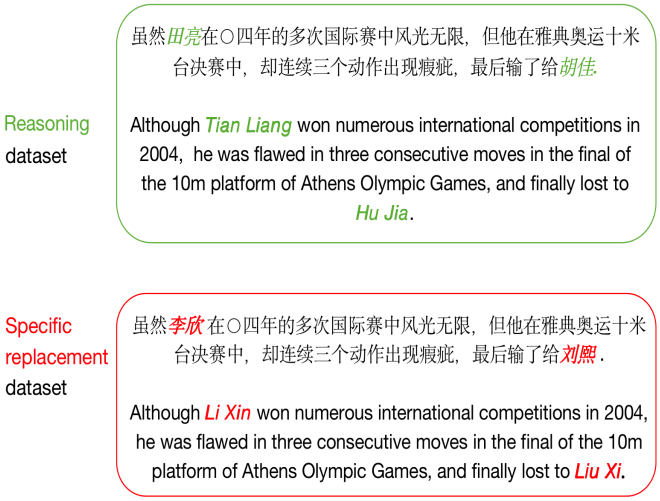

Fig. 3. Examples of reasoning instances specific replacement dataset.

The fifth specific replacement test dataset. This test dataset is base on the second reasoning test data, we replaced the specific entities(for example, a person's name, country/region, or other specific location) with another entity of the same level that has never appeared before. For example, we replace "田亮 TianLiang"and "刘熙 LiuXi" to "李欣 LiXin"and "胡佳 HuJia"in the sentence "虽然田亮在零四年的多次国际赛中风光无限，但他在雅典奥运十米台决赛中，却连续三个动作出现瘕疵，最后输了给胡佳 Although Tian Liang won numerous international competitions in 2004, he was flawed in three consecutive moves in the final of the 10m platform of Athens Olympic Games, and finally lost to Hu Jia". The entities "田亮 TianLiang"and "胡佳 Hu Jia"has appeared in previous instances, but the replaced words "李欣 LiXin"and "刘熙 Liu Xi"have never appeared before. The example is shown in Figure 3.

We replace the entities with other entities that have never appeared before, mainly to verify whether the BERT-based relation classification model will retain the information that learned before. Entities of the same level will not introduce additional information, we can still infer the relationship between entities through the meaning of the sentence if we have reasoning capability.

The final experiments show that the value of the F1-score on this test dataset is reduced compared to the second reasoning test dataset, it indicates our model will be impacted after changing the entities to the words that never appeared before. That is to say, BERT also uses the previous information to enrich the present representation for the same word.

C. BERT-based model

We proposed our BERT-based relation classification model based on Li's MG Lattice model [9]. Our model use BERT to enrich the word representation of the MG lattice model, and eventually exceed the MG Lattice model by 4.6% in F1-score. Same as the MG Lattice model, the input of our model is a Chinese sentence and a pair of entities. Figure 5 shows the overview of our model, which follows [9]. The more specific structure of our model will be introduced as follows.

Input Representation. The MG lattice model uses both characters-level and words-level information to enrich the word representations in the input sentence. We mainly combine BERT to enrich the characters-level representation of the MG lattice model. At the characters-level, given a sentence S which consists of M characters $S = \{c_1, c_2, \cdots, c_M\}$, we first use word2vec [4] same as MG lattice model, to map the character to the vector of d_e^c dimensions, and we got $x_i^{ce} \epsilon R^{d_e^c}$.

Besides, we use BERT [6] to enrich the character representation here, each of the character in the sentence $S = \{c_1, c_2, \cdots, c_M\}$, has mapped to the d_b^c via BERT. Then the output representation of BERT X_i^{cb} is fed into a fully connected layer that contains a linear activation. This layer will map the x_i^{cb} to d_e^c dimension. Now we get the character-lever representation x_i^{cb} via BERT.

In order to fully capture the information of the characters, we add x_i^{ce} and x_i^{cb} together to obtain more sufficient representation $x_i^{c'}$ for each of character in the sentence.

$$x_i^{c'} = x_i^{ce} + x_i^{cb} \tag{1}$$

For the final representation of the character, we also concatenate position representation at the character-level, which are defined as the relative distances from the current character to entity pairs. The position embedding of the i-th character c_i of two entities we represented as x_i^{p1},x_i^{p2} respectively. And finally we concatenate these three character representations to obtain the final characters-level representation x_i^c as shown below, which following [9]:

$$x_i^c = [x_i^{c'}, x_i^{p1}, x_i^{p2}] \tag{2}$$

The framework of the characters-level representation is shown in Figure 4.

The remaining structure of our model is the same as the MG Lattice. It will be briefly introduced below.

For Word-level representation, our model use lexicon D to match the word in input sentences to capture potential information from word-level. Then use Word2vec to map these words to vector representations $x_{b,e}^w$.

Since a word may have different senses, the model uses HowNet as the external knowledge base to find all senses for the word matched before, then also use word2vec to represent these senses information as vector representations $x_{b,e,1}^{sen}$, $x_{b,e,2}^{sen}$,...,$x_{b,e,n}^{sen}$.

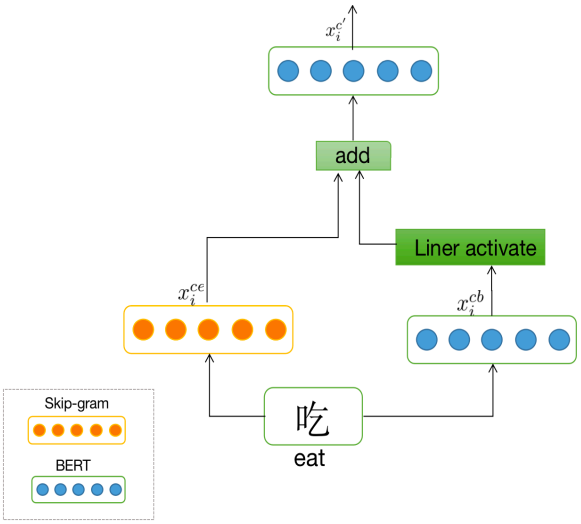

Fig. 4. Adding BERT to enrich the character level representations

Encoder The encoder of the model is improved based on LSTM. It takes word-level and character-level representations as the input. For the character-level representation, it can be used as the input to directly encoder the information with LSTM. But as the word-level representation, since different senses are introduced, a word may have several word-level representations. So the model

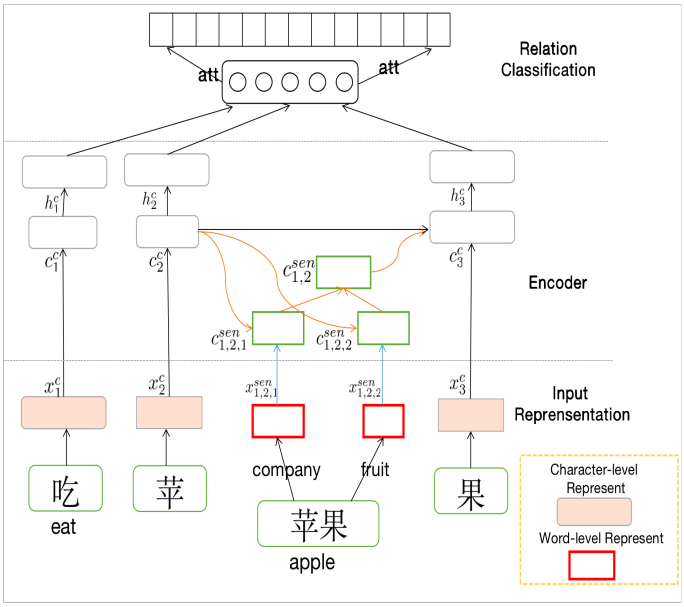

Fig. 5. The framework of BERT-based MG lattice model.

send each word-level representation to the basic lattice LSTM [24] for encoding to obtain the semantic cell state. Then multiply C_{sen} of the same word by the normalization factors α, and add them up to get the final semantic cell state of word C_{sen}. The model sends it to the next cell for calculation.

Relation Classifier This layer takes the hidden layer state which obtained from the encoder layer as input, and then adopt the attention mechanism to complete the final relation classification.

IV. Experiment

A. Dataset and Experimental Settings

Dataset. In this paper, we use the Chinese relation classification dataset from ACE Corpus 2005, which consists of 6 general relationships and 18 subtype relationships. We first manually annotate 800 reasoning instances that require a model to infer the relation between entity pairs based on the meaning of sentences. Then we select 600 reasoning instances as the first part of the training dataset, and the other 200 instances as the reasoning test dataset to evaluate the reasoning capability of our BERT-based model. For the remaining instances in ACE Corpus, we randomly select 80% as the second part of the training dataset and 20% instances as the relation classification test dataset to evaluate the performance of our BERT-based model in relation extraction task.

In addition, we create three other test datasets to comprehensively evaluate the reasoning capability of BERT and further explore how BERT enriches the word representation.

- Relation representation test dataset. We use “Entity A”and “Entity B”to masked the entities in the sen-

tences from reasoning dataset, to confirm whether BERT could infer the relationship of the entities in reasoning instances through the relation representation.

- Concept test dataset. This dataset based on the relation representation test dataset. We replace the masked entities with words such as "someone", "somewhere", which included conceptual information of the original entity. Compare with the relation representation test dataset, BERT can obtain more information from these entities.
- Specific replacement test dataset. This dataset based on the reasoning test dataset. We replace the specific entities with other entities of the same level which have never appeared before. It explores whether our BERT-based model can retain previously learned knowledge.

TABLE I
HYPER-PARAMETERS

Hyper-parameter	value
Character embedding size	100
Bert embedding size	768
LSTM hidden size	200
Position embedding size	5
Dropout probability	0.3
Learning rate	0.015

Experimental Settings. We use micro-averaged Precision(P),Recall(R) ,F1-score as the evaluation metric for our model. The SGD optimizer with the learning rate decay is utilized in our model training to tune the parameters. More detailed hyper-parameters of our model are shown in Table I.

B. Effect of BERT on the relation classification task.

In this part, we mainly focus on the effect of the BERT on the relation classification task. We evaluate our BERT-based MG lattice model on the first relation classification test dataset, and compare the model performance with various proposed deep learning network methods: CNN [17], BLSTM [2],and MG Lattice [9]. Table II shows the results.

From the table, we can see that the F1-score of our model reaches 82.9% , which outperforms all the other deep learning models. Especially compared with the MG lattice model, our BERT-based model improves the F1-scores of the first test datasets by 4.73 %. The results demonstrate that BERT will effectively improve the performance of the MG lattice model for the relation classification and it is reasonable to use our BERT-based relation classification model to explore the reasoning capability of BERT for the relation classification task.

C. Reasoning Capability of BERT

In this section, we will discuss the reasoning capability of BERT and further study what information BERT mainly uses to enrich word representations. Hence, we evaluate the precision(P), recall(R), and F1-score of our BERT-based model and MG lattice model on five different test datasets respectively. The results are shown in Table III.

TABLE II
THE F1-SCORE OF DIFFERENT DEEP LEARNING METHOD ON ACE CORPUS

Models	F1-score
CNN	72.41
BLSTM	70.69
MG Lattice	78.17
Our BERT-based MG Lattice	**82.9**

Reasoning Capability of BERT. First we notice the F1-score of our BERT-based model on the first test dataset and the second reasoning test dataset in Table III, has dropped from 82.9% to 79.67%. It demonstrates that the reasoning instance will impact the effectiveness of our BERT on the relation classification task. Nevertheless, we also notice that the F1-score of the MG lattice model has dropped from 78.17% to 65.75% in the same situation. Compare with the 13.6% decrease on MG lattice model, the performance of our BERT-based model is quite stable with only 3.2% reduction. Therefore we can conclude that BERT can greatly alleviate the impact than other pre-trained language models in the relation classification task.

Then we study whether BERT performs so well on these reasoning instances because of its reasoning capability. For the third relation representation instances, if BERT has reasoning capability, our model can still infer entity relationships based on the meaning of the sentence. However, the experimental results show that compare with the 79.67% F1-score on the reasoning dataset, the F1-score has dropped almost 39% on the relation representation dataset. BERT performs badly when it only uses the relation representation of the reasoning instances to predict the entities' relation, which indicates that BERT is not able to infer the relation between entities based on understanding the meaning of the sentences. In other words, BERT has no reasoning capability.

Information used by BERT. In this section, we mainly to explore how could our BERT-based model perform such excellent on the reasoning dataset even if it can not infer the relation between entities based on the meaning of sentences. From the Table III, we can find that compare to the poor F1-score on relation representation dataset, our model performs quite better on the concept test dataset with about 24.3% improvement. It shows that adding the conceptual information of entities will greatly improve the efficiency of the BERT-based model on reasoning instances. Especially, the MG Lattice model only has an 8.5% improvement on the F1 value in the same situation, which also confirms the excellent performance of BERT in learning contextual information of the sentences. Therefore, we can deduce that BERT mainly uses the

information of the entity to predict the relationship of the entity pairs instead of the meanings of sentences.

Finally we use the specific replacement dataset to discuss whether BERT can keep the knowledge that learned before. The F1-score of the specific replacement dataset has decreased by nearly 6.2% compared to the reasoning dataset, which means when entities are replaced with certain entities that have never appeared before, the performance of the BERT-based model will be impacted. So we can confirm that BERT also uses information learned from previously to enrich the current word representation and We think it is one of the reasons for Bert's outstanding performance in reasoning instances. Since the relationship between many entities also needs to be predicted by multiple sets of sentence level instances or document level information in real life. The function that BERT can retain the previously learned knowledge inspires further research on the relation classification task.

TABLE III
PERFORMANCE OF BERT-BASED AND MG LATTICE MODEL ON DIFFERENT DATASETS

Test Datasets	**MG Lattices**			**Bert-based model**		
	P	***R***	***F1***	***P***	***R***	***F1***
1	78.27	78.07	78.17	82.59	83.22	82.90
2	66.30	65.21	65.75	79.46	79.89	79.67
3	40.38	37.72	39.01	41.91	41.67	41.79
4	47.98	47.15	47.56	65.34	65.71	65.09
5	58.5	60.00	57.06	73.48	73.48	73.48

[a]1.Test dataset of whole RE model. 2.Reasoning dataset 3. Relation Representation dataset 4. Concept dataset 5. Specific Replacement dataset

Conclusion and Future work

In this paper, we propose a BERT-based relation extraction model and five different kinds of test datasets to explore the reasoning capability of BERT on the relation classification task. We first use BERT to enrich the character-level representation of the MG lattice model to train our BERT-based model and it achieves a quite improvement than the original MG lattice model on ACE Corpus. Through various comparison experiments on five test datasets, we found that although our BERT-based model performs excellent than MG lattice model in each test dataset, especially in reasoning instances, it still can not infer the relationships between entity pairs based on understanding the meaning of the sentences.

Further, form the results of the experiments we conclude that BERT perform better on reasoning instances because it not only uses the concept information of the entities itself, but also keeps the features of entities that learned before to enrich the word representations. And these characteristics inspire us to use BERT to further study the relation extraction when the relationship between entities is defined by multiple sets of sentences.

In future work, we plan to work on combining the graph neural networks such as GCN with our BERT-based MG lattice model to further improve the power of our relation extraction model on Chinese corpus. It is worth noting that most of the existing GCN-based relation classification models are built on English corpora. Therefore, the combination of graph neural network and BERT can deeply study how to improve the efficiency of the neural network in the relation extraction of Chinese reasoning instances.

Acknowledgement

This work is supported by Beijing Natural Science Foundation (No. 4202069) and National Natural Science Foundation of China (No.61672100).

References

[1] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic Compositionality through Recursive Matrix-Vector Spaces," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

[2] Z. Peng, S. Wei, J. Tian, Z. Qi, and X. Bo, "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification ," in *Meeting of the Association for Computational Linguistics*, 2016.

[3] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural Relation Extraction with Selective Attention over Instances," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrasesand their Compositionality," *Advances in Neural Information Processing Systems*, 2013.

[5] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, vol. abs/1810.04805, 2018.

[7] S. Wu and Y. He, "Enriching Pre-trained Language Model with Entity Information for Relation Classification," *CoRR*, vol. abs/1905.08284, 2019.

[8] L. B. Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski, "Matching the Blanks: Distributional Similarity for Relation Learning," *CoRR*, vol. abs/1906.03158, 2019.

[9] Z. Li, N. Ding, Z. Liu, H. Zheng, and Y. Shen, "Chinese Relation Extraction with Multi-Grained Information and External Linguistic Knowledge," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[10] N. Kambhatla, "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, (Barcelona, Spain), pp. 178–181, Association for Computational Linguistics, July 2004.

[11] C. Giuliano, A. Lavelli, D. Pighin, and L. Romano, "FBK-IRST: Kernel Methods for Semantic Relation Extraction," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, (USA), p. 141 – 144, Association for Computational Linguistics, 2007.

[12] S. Tratz and E. Hovy, "ISI: Automatic Classification of Relations between Nominals Using a Maximum Entropy Classifier," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, (USA), p. 222 – 225, Association for Computational Linguistics, 2010.

[13] A. Culotta, A. McCallum, and J. Betz, "Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, (USA), p. 296–303, Association for Computational Linguistics, 2006.

[14] D. Zelenko, C. Aone, and A. Richardella, "Kernel Methods for Relation Extraction," *J. Mach. Learn. Res.*, vol. 3, p. 1083 – 1106, Mar. 2003.

[15] S. Zhao and R. Grishman, "Extracting Relations with Integrated Information Using Kernel Methods," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, (USA), p. 419 – 426, Association for Computational Linguistics, 2005.

[16] M. Zhang, J. Zhang, and J. Su, "Exploring Syntactic Features for Relation Extraction Using a Convolution Tree Kernel," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, (USA), p. 288–295, Association for Computational Linguistics, 2006.

[17] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," *the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2335–2344, 01 2014.

[18] W. Ye, B. Li, R. Xie, Z. Sheng, L. Chen, and S. Zhang, "Exploiting Entity BIO Tag Embeddings and Multi-task Learning for Relation Extraction with Imbalanced Datasss," *CoRR*, vol. abs/1906.08931, 2019.

[19] L. Zhang and D. Moldovan, "Chinese Relation Classification using Long Short Term Memory Networks," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, eds.), (Miyazaki, Japan), European Language Resources Association (ELRA), May 7-12, 2018 2018.

[20] A. Santoro, D. Raposo, D. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," *neural information processing systems*, 06 2017.

[21] P. Battaglia, J. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, and R. Pascanu, "Relational inductive biases, deep learning, and graph networks," *arXiv: Learning*, 2018.

[22] Y. Zhang, P. Qi, and C. D. Manning, "Graph Convolution over Pruned Dependency Trees Improves Relation Extraction," *CoRR*, vol. abs/1809.10185, 2018.

[23] Z. Guo, Y. Zhang, and W. Lu, "Attention Guided Graph Convolutional Networks for Relation Extraction," *CoRR*, vol. abs/1906.07510, 2019.

[24] Y. Zhang and J. Yang, "Chinese NER Using Lattice LSTM," *meeting of the association for computational linguistics*, vol. 1, pp. 1554–1564, 2018.