IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Novel Document-Level Relation Extraction Method Based on BERT and Entity Information

XIAOYU HAN[1,2,3], LEI WANG[1,2]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[2]Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[3]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Lei Wang(e-mail: wanglei@mail.ie.ac.cn).

ABSTRACT Document-level relation extraction aims to extract the relationship among the entities in a paragraph of text. Compared with sentence-level, the text in document-level relation extraction is much longer and contains many more entities. It makes the document-level relation extraction a harder task. The number and complexity of entities make it necessary to provide enough information about the entities for the models in document-level relation extraction. To solve this problem, we put forward a document-level entity mask method with type information (DEMMT), which masks each mention of the entities by special tokens. By using this entity mask method, the model can accurately obtain every mention and type of the entities. Based on DEMMT, we propose a BERT-based one-pass model, through which we can predict the relationships among the entities by processing the text once. We test the proposed model on the DocRED dataset, which is a large scale open-domain document-level relation extraction dataset. The results on the manually annotated part of DocRED show that our approach obtains 6% F1 improvement compared with the state-of-the-art models that do not use pre-trained models and has 2% F1 improvement than BERT which does not use the DEMMT. On the distant supervision generated part of DocRED, the improvement of F1 is 2% compared with no pre-trained models, and 5% compared with pure BERT.

INDEX TERMS Relation extraction, document-level, BERT, entity information, one-pass.

I. INTRODUCTION

RELATION extraction aims to assign semantic relationships to the entities in the unstructured text which is a vital part in many natural language applications, such as information extraction, question answering, and knowledge graphs construction. According to the scope of text processing, relation extraction can be divided into sentence-level relation extraction and document-level relation extraction. The goal of sentence-level relation extraction is to obtain the relation between two known entities in a sentence. While document-level relation extraction aims to extract the relationship among several entities in a long text that usually contains multiple sentences, as shown in Figure 1.

Recently, most of the related works are based on the idea of sentence-level. However, only focus on the relation of entities in sentence-level is not enough. Previous work [1] shows that, in MUC6 (Message Understanding Conference) [2] and ACE03 (Automatic Content Extraction) corpus, there are 28.5% and 9.4% of inter-sentential relations respectively, whose relationships between two entities expressed over more than one sentence, among the total number of relations. The DocRed [3] is a manually annotated dataset generated from Wikipedia, in which there are over 40.7% relational facts can only be extracted from multiple sentences. The above statistics reveal that the amount of inter-sentential relations is too large to be ignored. Therefore, it is necessary and important to study document-level relation extraction.

There are giant differences between the schemes of document-level and sentence-level relation extraction, some major of them are listed as the following:

(1) In document-level relation extraction, the entity number is much larger than that of sentence-level, therefore, all the relationships among the different entities need to be considered. For example, we need to judge the relations between 20 entity pairs in Figure 1.

(2) In sentence-level relation extraction, an entity usually

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2996642, IEEE Access

Document

Anzac biscuit

[1] An Anzac biscuit is a sweet biscuit, popular in Australia and New Zealand, made using rolled oats, flour, sugar, butter (or margarine), golden syrup, baking soda, boiling water, and (optionally) desiccated coconut. [2] Anzac biscuits have long been associated with the Australian and New Zealand Army Corps (ANZAC) established in World War I. [3] The biscuits were sent by wives and women's groups to soldiers abroad because the ingredients do not spoil easily and the biscuits kept well during naval transportation. [4] Today, Anzac biscuits are manufactured commercially for retail sale. [5] Anzac biscuits should not be confused with hardtack, which was nicknamed "ANZAC wafers" in Australia and New Zealand.

Triples

Subject: Anzac biscuit	Object: Australia	Relation: country	Supporting Evidence: 1, 5
Subject: Anzac biscuit	Object: ANZAC	Relation: associated_with	Supporting Evidence: 2
Subject: ANZAC	Object: World War I	Relation: established_in	Supporting Evidence: 2

Figure 1. Example of document-level relation extraction.

occurs once in the sentence. While an entity may appear several times in different forms in document-level relation extraction, for example, the entity 'Anzac biscuits' is also represented as 'ANZAC wafers' in Figure 1.

(3) In document-level relation extraction, the relationship between two entities may not be extracted directly. The transitivity of relations must be considered in order to find some relations.

According to the differences mentioned above, document-level relation extraction is more complicated than that of sentence-level. In this paper, we concentrate on the document-level relation extraction, a more challenging task than traditional sentence-level one.

Many methods have been proposed for the relation extraction task in the past years, including the traditional methods [4], [5] which rely on manual feature engineering and the neural-network-based models [6]–[8] which extract features by end-to-end training and achieve the state-of-the-art performance. These neural-network-based methods utilize position feature, which gives the relative distance to the two entities of each word as an input of the models, to obtain entity information. Some recent works apply pre-trained models, such as BERT [9], into relation extraction. Wang et al. [10] simply utilizes the pre-trained transformers [11] to encode the text into the contextual representations, and achieves about 3 points higher on F1-Measurement. The position feature is not applicable in the pre-trained models because of the input constraint. Therefore, the information of the target entities is missing during the encoding of the text. It is believed that providing the entity position to the pre-trained models would further improve the performance. In sentence-level, Wu et al. [12] propose to use special tokens to mark the two entities in the sentence which can be used in pre-trained models, and it achieves great progress. As there are multiple target entities in document-level relation extraction, these entity mask methods, which give information of two entities in one time, is no longer qualified for they can not provide information of all the entities in one time. To solve this problem, Yao et al. [3] use the NER information and entity index feature instead of the position feature to mark the entities. However, like the position feature, these features can not be used in pre-trained models.

In this paper, we extend the approach which uses special tokens to mark the entities in sentence-level relation extraction. In our method, we use different tokens to mark different entities, as well as provide the NER type of each entity, so the models can get enough information about the entities. A BERT-based one-pass model is proposed based on the entity mask method. By using the entity mask method, the model can handle these difficulties of document-level relation extraction. First, unlike the position feature which can only deal with two entities, the proposed entity mask method can provide information of all the entities no matter how many entities there exists. Therefore, the one-pass model can predict the relationship between all the entity pairs by only encoding the text once. Second, the proposed mask method label the same entity with the same mask so that no matter how many forms an entity exists the model can recognize that they refer to the same entity. Last, the one-pass method used in our method can utilize the information of every entity which is helpful to handle the transitivity of relations. The experiment result shows that the proposed model outperforms the current state-of-the-art models.

The contributions of this paper are as follows:

First, we put forward an entity mask method which can introduce entity identity and type information to the models.

Second, we proposed a BERT-based one-pass model which introduces the entity information by using the proposed entity mask method, and the proposed model achieves state-of-the-art performance on the DocRed dataset [3].

The remainder of this paper is structured as follows. Section II lists recent works related to document-level relation extraction. Section III presents the proposed methods in detail. Section IV provides the dataset and experimental results. The conclusion is illustrated in section V.

II. RELATED WORK

With the development of deep learning, neural-network-based models become more and more popular in many areas [13], [14] including relation extraction. Recursive Neural Network was first used in relation extraction by Socher et al. [6] who propose the Recursive Matrix-Vector Model to model the SDP (shortest dependency path) between entities in the sentence. CNN (Convolution Neural Network) was introduced into the relation extraction task by Zeng et al. [7]. Subsequently, Zhang et al. [8] selected RNN (Recurrent Neural Network) to model the sentences directly. Based on these methods, many improved networks are proposed, such as PCNN [15] and BRCNN [16], the former uses piece-wise max-pooling in CNN and the later combines the RNN and CNN. Then, as the effect of attention mechanism is proved in NLP tasks, this mechanism is widely used in relation extraction. Zhou et al. [17] proposed an attention-based LSTM (Long Short-term Memory Networks) model. Geng et al. [18] apply attention-based bidirectional tree-structured LSTM to extract structural features based on the dependency tree of a sentence. Wang et al. [19] proposed a model that uses multi-level attention CNNs. In addition, GCN (Graph Convolutional Network) [20] has become popular in the NLP area in recent years. Zhang et al. [21] use GCN over the pruned dependency trees to improves relation extraction. Guo et al. [22] proposed AGGCNs which use the attention mechanism on GCN to find the important edges of the dependency trees.

In all the methods mentioned above, word embeddings [23], [24] are used to represent the words. Another way of contextual representation of each word is to employ the pre-trained language models [9], [25], [26]. Devlin et al. [9] propose to learn the contextual representations of each word by training a large scale language model. They proposed the BERT and broken the record of many NLP tasks, such as the text classification, NER (Named Entity Recognition), text similarity measurement, and reading comprehension. Shi et al. [27] applied a simple BERT-based model for both relation extraction and semantic role labeling. Wu et al. [12] improved the BERT model for relation classification by inserting special tokens before and after the target entities before feeding the text to BERT for fine-tuning. Wang et al. [28] extracted multiple-relations in one-pass with BERT.

Compare with sentence-level relation extraction, document-level relation extraction can extraction inter-sentential relationships between entities. Swampillai et al. [1] proved that inter-sentential relation made up a relatively significant share of all the relations by analyzing two corpora. Quirk et al. [29] expanded the dependency trees to get the graph feature to increase the accuracy and robustness of inter-sentential relation extraction. Peng et al. [30] explored a graph LSTM based framework that was capable for cross-sentence n-ary relation extraction. Wang et al. [28] used Transformer [11] to extract the relationship between all entity pairs in one-cross. Yao et al. [3] made DocRED, a large-scale human-annotated document-level relation extraction dataset constructed from Wikipedia and Wikidata, and test several state-of-the-art neural network models on it. Wang et al. [10] fine-tuned BERT with a two-step process on the DocRED and achieved state-of-the-art performance.

III. PROPOSED MODEL

In this part, we describe the proposed approach for document-level relation extraction in detail. The overall structure of the proposed approach is shown in Figure 2. There are three compositions of the proposed approach. The first one is the entity mask method, which provides the entity information for the other parts. The second one is the pre-train model BERT, which is used to encode the text. The last part is the classification neural-network, which predicts the relationship of all the entity pairs in one-pass.

A. DOCUMENT-LEVEL ENTITY MASK METHOD WITH TYPE INFORMATION

In document-level relation extraction, the number of entities is much larger than that of sentence-level. Methods widely used in sentence-level which mark two entities in one time are not qualified. In document-level, we need an entity mask method that can clearly describe every mention of each entity. So, we propose a document-level entity mask method with type information (DEMMT) which is shown in Figure 3.

In the proposed method, each mention of an entity is masked by two tokens. The first token is before the first word of the mention, which represents the NER type of the entity. In this paper, we use six entity types that are widely used in the NER task, including ORG (Organization), LOC (Location), TIME (Time), PER (Person), NUM (Number), and MISC (miscellaneous). By using these entity types of each entity, we can provide the high-level semantic information to the model. The second token is after the last word of the mention, which represents the entity that the mention linked to. In document-level relation extraction, each entity may be mentioned in different forms several times. Entity identity that points out which entity the mention is described is necessary. For example, in Figure 3, there are five entities and marked by 'MASK_1' to 'MASK_5'. We take the first entity, 'Anzac biscuit', as an example. The NER type of the entity is 'MISC'. The entity has two forms which are 'Anzac biscuit' and 'ANZAC wafers', and there exist four mentions in the text. Each mention is represented as '[MISC] Anzac biscuit [MASK_1]' or '[MISC] ANZAC wafers [MASK_1]' according to the form.

In document-level relation extraction, an entity mask method has to satisfy two essential factors. First, every mention of the entities in the text must be marked. Second, which entity is the mentions linked to must be provided. The proposed DEMMT method satisfies the two factors above. Besides, it also provides NER type information for the model which is helpful for relation extraction.

When conducting the mask method on BERT, unlike other word embeddings based neural-network which can randomly generate a vector to represent the mask, each input of the BERT model must be a token that is known by BERT.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2996642, IEEE Access

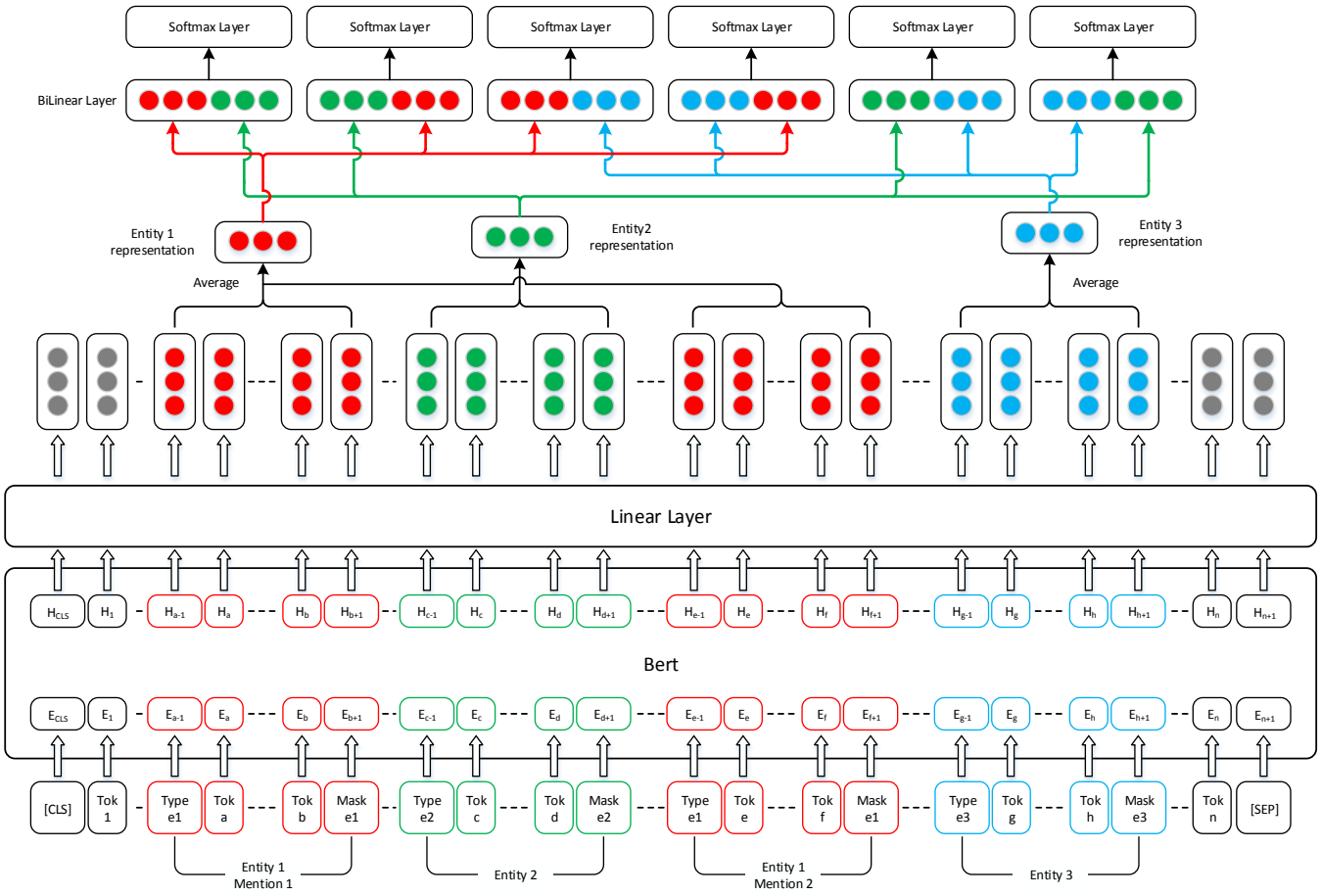

Figure 2. The proposed model.

[1] An [MISC] Anzac biscuit [MASK_1] is a sweet biscuit, popular in [LOC] Australia [MASK_2] and [LOC] New Zealand [MASK_3], made using rolled oats, flour, sugar, butter (or margarine), golden syrup, baking soda, boiling water, and (optionally) desiccated coconut. [2] [MISC] Anzac biscuits [MASK_1] have long been associated with the [ORG] Australian and New Zealand Army Corps [MASK_4] ([ORG] ANZAC [MASK_4]) established in [MISC] World War I [MASK_5]. [3] The biscuits were sent by wives and women's groups to soldiers abroad because the ingredients do not spoil easily and the biscuits kept well during naval transportation. [4] Today, [MISC] Anzac biscuits [MASK_1] are manufactured commercially for retail sale. [5] [MISC] Anzac biscuits [MASK_1] should not be confused with hardtack, which was nicknamed "[MISC] ANZAC wafers [MASK_1]" in [LOC] Australia [MASK_2] and [LOC] New Zealand [MASK_3].

Figure 3. Example of document-level entity mask method with type information.

These masks will be replaced as unknown tokens that will make these masks dysfunctional. Therefore, we use special tokens, which are rarely used in English text but is known by BERT, to replace these masks. For example, we use 'æ' to replace the mask '[MISC]' and '¡' to replace '[MASK_1]', the mention 'Anzac biscuit' is marked as 'æAnzac biscuit ¡' after replacement.

B. BERT ENCODER

BERT (Bidirectional Encoder Representation from Transformers) is a large pre-trained language model proposed by Google [9] in 2018. Since then, BERT has achieved state-of-the-art results on various NLP tasks. BERT is a powerful pre-trained model which is consists of multi-layer bidirectional Transformer encoder [11]. There are two steps in the BERT framework, namely the pre-training and fine-tuning. By the pre-training step, BERT can obtain abundant background information about the target language. The information is stored by hundreds of millions of parameters in BERT. In the fine-tuning step, BERT is initialized with the pre-trained parameters obtained in the pre-training step. Then, the parameters are fine-tuned by the labeled data of the downstream task.

In our approach, we use BERT to encode the text that in-

cludes the entities which we want to extract the relationships among them. The parameter of BERT is initialized by the parameters which are provided by Google which is trained on BooksCorpus and English Wikipedia. The text is first marked by the proposed entity mask method to provide entity information for BERT.

The input is a piece of text $D = \{T_1, T_2, \ldots, T_n\}$, in which the entities are masked by the proposed entity mask method. The i-th entity in the text is represent as $E_i = \{E_i^1, \ldots, E_i^k\}$ where E_i^k is the k-th mention of the entity, and $E_i^k = \{T_a, \ldots, T_b\}$ where T_a to T_b is the tokens in the mention including the two mask tokens. Before putting into the BERT model, the token $[CLS]$ and $[SEP]$ are added at the head and end of the sequence respectively. The encoding of the tokens $H = \{H_1, H_2, \ldots, H_n\}$ is calculated as follows:

$$\{H_1, H_2, \cdots, H_n\} = BERT(\{T_1, T_2, \ldots, T_n\}) \quad (1)$$

C. ONE-PASS RELATION PREDICTION

In document-level relation extraction, the relationship is needed to be predicted between all the entity pairs that appear in the input text. Consider there are k entities in the text, then we have to predict the relationship of $k \times (k-1)$ entity pairs if we have to discriminate the direction of the relationship. In document-level relation extraction, the number of entities is generally far more than two, in some cases, it can even reach dozens. This means that we may need to deal with hundreds of entity pairs. If we need to model sentences every time we process entity pairs, it will cause a huge amount of computation, which is not acceptable. And this is also the reason why the position feature, which is widely used in sentence-level relation extraction to mark the entities, is not capable of document-level. So, we use a one-pass relation prediction method to process all the entity pairs in one time.

After getting the hidden states H calculated by BERT, we use a linear layer to get the embedding h_i of T_i. The linear layer is a full connection layer that is used to extract key information from the BERT outputs.

$$h_i = Linear(H_i) \quad (2)$$

Then we need to get the representations of the entities. The representation of an entity is calculated as the average of all the mentions of the entity. And the representation of a mention is the average of all the embeddings of the tokens in it. The calculated process is as follows.

$$e_i = \frac{1}{j}\sum_{k=1}^{j}(e_i^k) \quad (3)$$

In the equation, e_i is the embedding of entity i, e_i^k is k-th mention of entity e_i, and j is the number of mentions of entity e_i. The embedding of mention e_i^k is calculated as follows, where $h_l^{e_i^k}$ is the embedding of the l-th token, m is the number to tokens in the mention.

$$e_i^k = \frac{1}{m}\sum_{l=1}^{m}(h_l^{e_i^k}) \quad (4)$$

Then we use a bilinear layer, which is a full connection layer using two inputs, to get the representation of an entity pair. Entity pair (e_i, e_j) is represented as:

$$R_{(e_i,e_j)} = BiLinear(e_i, e_j) \quad (5)$$

At last, a softmax layer is used to predict the relations of each entity pairs. In this way, we can obtain the relations of all entity pairs under the condition of only encoding the text once.

IV. EXPERIMENTS

In this part, a series of experiments are performed to test the proposed document-level relation extraction method. First, we compared our method with previous state-of-the-art methods. After that, we try to find out how the proposed DEMMT work and in which way we can take the most advantages of the entity information through a set of experiments.

A. DATA SELECTION

In this paper, we select the DocRED dataset [3], which is a document-level relation extraction data generated by distant supervision [31], to verify the proposed approach. DocRED used Wiki-data, a large-scale KB tightly integrated with Wikipedia, to label the introductory sections from English Wikipedia. The labeled data includes NER result, entity linking result, and relation information, that can be used in future works. After generating the data with distant supervision, part of the data is select for human annotation. So, DocRED has two parts of constitutions, one is the distant supervision generated part which contains 101,873 documents, and the other one is the manually annotated part which includes 5053 documents. The manually annotated part is divided into the training part, develop part, and test part. The instance number of the training part is 3053 and the instance number of develop part and test part are both 1000.

The coverage of the DocRED dataset is very wide. There exist 96 frequent relation types from Wikidata, these relations are relevant to science, art, personal life, and so on. Most relation instances in the dataset need reasoning, even logical reasoning, to be identified. The proportion of inter-sentential relation fact is 46.4%, and 40.7% relational facts can only be extracted from multiple sentences. So, the DocRED dataset is a large scale open-domain relation extraction dataset that contains lots of inter-sentential relation facts. It is very suitable for estimating document-level relation extraction approaches.

B. IMPLEMENTATION DETAILS

There are several versions of BERT, of which the BERT-Base (uncased) is selected as our encoder in the approach. The BERT-Base model contains 12-layers of Transformer,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2996642, IEEE Access

and each Transformer uses 12-heads self-attention. The size of hidden states in BERT-Base is set to 768. Other related parameters are list in Table 1.

Table 1. Parameter Settings.

Parameter	Value
Batch Size	4
Max Sentence Length	512
Learning Rate	10e-5
Entity Embedding Size	128
Entity Pair Embedding Size	97
Positive/Negative Rate	1/6

The last row in Table 1 is the ratio of the positive examples and negative examples in the training process. As mentioned above, given a document contains k entities, we can have $k \times (k-1)$ entity pairs. Only a small part of these entity pairs have specific relationships which can be seen as positive examples. If we use all the entity pairs in the training process, this will lead to a serious imbalance between positive and negative examples. Therefore, we randomly resample negative samples in a designative proportion.

C. COMPARISION WITH OTHER MODELS

In this section, we compare the proposed approach with several baseline models including state-of-the-art neural-network in sentence-level relation extraction, a model that is designed for leveraging contextual relations which can improve intra-sentence relation extraction, and models based on pre-trained language models. These model includes a CNN-based model [7], a LSTM-based model [32], a BiLSTM based model [16], the Context-Aware model [33], and a BERT-based model [10]. The first three models use different encoders to encode the documents. The Context-Aware model considers the sentential context using the attention mechanism. And the BERT-based model introduces BERT to document-level relation extraction. The models are tested on both the annotated data and the distant supervision generated data of the DocRED dataset. Tables 2 shows the performance of these models.

The 'Ign F1' in Table 2 is the F1 score that excludes the entity pair that appears in the training data. From this table, we can conclude that the proposed BERT+DEMMT model outperforms other baseline models by a significant margin in the F1 score. Among all these models, the CNN-based model has the lowest F1 score in all the situations. The reason might be that the shallow CNN is inappropriate to encode long text sequences. Compared with CNN, LSTM and BiLSTM can deal with long text sequence better. So, the LSTM and BiLSTM based models achieve much better results than the CNN-based one. To our surprise, the Context-Aware model that is designed to utilize the interaction between entity pairs just show similar performance with the BiLSTM-based model. By using the pre-trained language model, BERT has a head start over other models because of the knowledge about language. But, it seems more vulnerable to the missing or wrong labels in the data that is introduced by distant supervision. The proposed BERT+DEMMT model gets the new stat-of-the-art result by providing BERT enough entity information. On the manually annotated data of DocRED, it obtains a 6% F1 improvement compared with the first four models and has a 2% F1 improvement compared with BERT which does not use the DEMMT. From the result of the distant supervision generated data, the DEMMT greatly helps to narrow down the problem caused by the missing and wrong labels in BERT.

D. ABLATION EXPERIMENTS

In this section, we aim to find out how the proposed DEMMT work and in which way we can take the most advantages of the entity information. All the experiment in this section is implemented on the development set of manually annotated data in DocRED.

1) Entity Mark Method

In this part, we compare four entity mask methods to find out the effectiveness of DEMMT. The methods are as follows:

The first one is BERT, which does not use any masks, in this situation the model is the same as Wang et al. used [10].

The second one is BERT+TT, which only uses NER types to mark each entity. The entity mentions in BERT-TT are masked as '[TYPE] Entity Mention [TYPE]'.

The third one is BERT+MM, which selects the entity mask to label the entities. The entity mentions in BERT-MM are masked as '[MASK_IDX] Entity Mention [MASK_IDX]', where the 'IDX' is the index of the entity in the document.

The last one is BERT+DEMMT, which is the proposed approach.

The results are shown in Table 3. According to this table, we can know that any kind of mask can improve the result. The improvement of BERT+TT is not significant compared with BERT+MM. By BERT+TT, the BERT model reveals which word in the document represents an entity and what type is the entity related to. By BERT+MM, the model can not only know which word represents an entity but also know which entity each mention is linked to. Therefore, we find that the operation of linking the mentions to the specific entity is more helpful than only provide the type information. However, from the gap between BERT+DEMMT and BERT+MM, we can say that the NER type information is more than just point the entities.

2) Entity Representation

In the proposed method, each mention is represented as the average of all the token embedding including the masks. In this section, we will illustrate if the embedding of the masks carries important information, or the information is already got by the embedding of the words in the mention.

We test the following four experiments to confirm the assumption. The first experiment uses the average of the words embedding in the mentions, which do not include the masks. The second one and third one utilize the type mask

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2996642, IEEE Access

Table 2. Performance of different models on DocRED.

Model	Manually Annotated Data				Distant Supervision Generated Data			
	Dev		Test		Dev		Test	
	Ign F1	F1	Ign F1	F1	Ign F1	F1	Ign F1	F1
CNN	41.58	43.45	40.33	42.26	33.24	42.76	32.33	42.00
LSTM	48.44	50.68	47.71	50.07	39.37	49.92	38.27	48.88
BiLSTM	48.87	50.94	48.78	51.06	41.44	51.72	39.15	49.80
Context-Aware	48.94	50.17	48.40	50.70	40.47	51.39	39.16	50.12
BERT	53.35	55.15	52.53	54.83	37.35	48.68	35.40	47.16
BERT+DEMMT(Proposed)	55.50	57.38	54.93	57.13	42.56	53.33	41.76	52.45

Table 3. Entity mask method comparison.

Method	Ign F1	F1
BERT	53.35	55.15
BERT+TT	53.19	55.36
BERT+MM	55.13	57.19
BERT+DEMMT	55.50	57.38

and entity mask with the words respectively. The last one is the proposed method that employs all the masks.

Table 4. Entity representation method comparison.

Method	Ign F1	F1
Word	54.64	56.88
Word+TypeMask	54.76	57.09
Word+EntityMask	54.95	57.22
Word+TypeMask+EntityMask	55.50	57.38

The result shows in Table 4 demonstrated that the embedding of the masks can provide information that does not be contained in the words. What consistent with the previous part is the information provided by the NER type mask has fewer impacts on the results than the entity mask. Besides, by using the masks, the embeddings of the words that consist of the mention can learn more things than no mask.

3) Application Environment of DEMMT

The effectiveness of DEMMT on BERT has been verified by the experiments above. In this section, we test the DEMMT to see whether it can bring improvement to other encoders. We adopt the encoders mentioned above which are CNN, LSTM, BiLSTM, and the Context-Aware model. The result is shown in Table 5

Table 5. Test DEMMT On Other Encoders.

Encoder	No DEMMT		With DEMMT	
	Ign F1	F1	Ign F1	F1
CNN	41.58	43.45	45.14	47.22
LSTM	48.44	50.68	47.34	49.26
BiLSTM	48.87	50.94	47.51	49.48
Context-Aware	48.94	50.17	48.54	50.34

From Table 5, we can conclude that DEMMT can bring giant benefits to CNN-based models. But the LSTM-based models, including the Context-Aware model, can not have a good handle on the way of introducing entity information in DEMMT, because that the CNN encoder can focus on the local features. So, it can notice the information provided by the masks which are around the entities. The LSTM encoder may lace a higher premium on the global features so that the information provided by the masks are ignored. Therefore, we obtain the conclusion that the DEMMT can achieve better performance on CNN encoders and Transformers encoders, which employ self-attention mechanism, than LSTM encoders.

V. CONCLUSION

Document-level relation extraction is a challenging and vital task in NLP. Compared with sentence-level relation extraction, the text in document-level relation extraction is much longer and contains more entities. In this paper, we propose a document-level entity mask method DEMMT to provide information about the entities in the text. The DEMMT method uses special tokens to mark the entities to provide NER type and entity identity information. To verify the effectiveness of DEMMT, we propose the BERT+DEMMT models which adopt the DEMMT with BERT encoder. The results demonstrate the BERT+DEMMT model achieves state-of-the-art performance compared with previous methods. Besides, we also conduct a series of experiments to explain how the DEMMT works and in which situation we can use DEMMT.

REFERENCES

[1] K. Swampillai and M. Stevenson, "Inter-sentential relations in information extraction corpora." in LREC, 2010.
[2] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996.
[3] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun, "Docred: A large-scale document-level relation extraction dataset," arXiv preprint arXiv:1906.06127, 2019.
[4] Z. GuoDong, S. Jian, Z. Jie, and Z. Min, "Exploring various knowledge in relation extraction," in Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005, pp. 427–434.
[5] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 455–465.
[6] 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. Association for Computational Linguistics, 2012, pp. 1201–1211.

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

[8] D. Zhang and D. Wang, "Relation classification via recurrent neural network," arXiv preprint arXiv:1508.01006, 2015.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[10] H. Wang, C. Focke, R. Sylvester, N. Mishra, and W. Wang, "Fine-tune bert for docred with two-step process," arXiv preprint arXiv:1909.11898, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[12] S. Wu and Y. He, "Enriching pre-trained language model with entity information for relation classification," in Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2361–2364.

[13] Z. Geng, Z. Li, and Y. Han, "A new deep belief network based on rbm with glial chains," Information Sciences, vol. 463, pp. 294–306, 2018.

[14] Y. Han, S. Zhang, Z. Geng, Q. Wei, and Z. Ouyang, "Level set based shape prior and deep learning for image segmentation," IET Image Processing, vol. 14, no. 1, pp. 183–191, 2019.

[15] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1753–1762.

[16] R. Cai, X. Zhang, and H. Wang, "Bidirectional recurrent convolutional neural network for relation classification," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 756–765.

[17] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 207–212.

[18] Z. Geng, G. Chen, Y. Han, G. Lu, and F. Li, "Semantic relation extraction using sequential and tree-structured lstm with attention," Information Sciences, vol. 509, pp. 183–192, 2020.

[19] L. Wang, Z. Cao, G. de Melo, and Z. Liu, "Relation classification via multi-level attention cnns," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1298–1307.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[21] Y. Zhang, P. Qi, and C. D. Manning, "Graph convolution over pruned dependency trees improves relation extraction," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2205–2215.

[22] Z. Guo, Y. Zhang, and W. Lu, "Attention guided graph convolutional networks for relation extraction," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 241–251.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111–3119.

[24] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in Advances in neural information processing systems, 2019, pp. 5754–5764.

[26] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, "Knowledge enhanced contextual word representations," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 43–54.

[27] P. Shi and J. Lin, "Simple bert models for relation extraction and semantic role labeling," arXiv preprint arXiv:1904.05255, 2019.

[28] H. Wang, M. Tan, M. Yu, S. Chang, D. Wang, K. Xu, X. Guo, and S. Potdar, "Extracting multiple-relations in one-pass with pre-trained transformers," arXiv preprint arXiv:1902.01030, 2019.

[29] C. Quirk and H. Poon, "Distant supervision for relation extraction beyond the sentence boundary," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 1171–1182.

[30] N. Peng, H. Poon, C. Quirk, K. Toutanova, and W.-t. Yih, "Cross-sentence n-ary relation extraction with graph lstms," Transactions of the Association for Computational Linguistics, vol. 5, pp. 101–115, 2017.

[31] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009, pp. 1003–1011.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] D. Sorokin and I. Gurevych, "Context-aware representations for knowledge base relation extraction," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1784–1789.

XIAOYU HAN received the B.Sc. degree from Harbin University of Science and Technology, Harbin, China, in 2011. He received the M.Sc degree from Guilin University of Electronic Technology, Guilin, China, in 2014. He is currently pursuing the Ph.D. degree with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include natural language processing, deep learning, especially on relation extraction and knowledge graph construction.

LEI WANG received the B.Sc. and M.Sc. degrees from Shandong University, in 1995 and 1998 respectively. He received the Ph.d. degree from Institute of Electronics, Chinese Academy of Science, in 2001.

He is a currently Professor with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include data mining, large scale data organization technology, and geospatial information application technology.