Preprints.org

Article

ETFRE: Entity Type Fusing for Relation Extraction

Peilin Shi , Bin Zhang , Yingkun Liu , Cheng Fang [*]

Article

ETFRE: Entity Type Fusing for Relation Extraction

Peilin Shi [1], Bin Zhang [1], Yingkun Liu [2] and Cheng Fang [2,*]

[1] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China
*   Correspondence: cfang@cauc.edu.cn

Abstract: This paper proposes a relational extraction framework based on entity type information fusion by Transformer model. Relational extraction, as an important part of knowledge graph construction, has been paid much attention in recent years. The existing relational extraction and joint triple extraction models rarely use the existing entity type information, so the semantic features of the entity type are lost, resulting in limited model performance and difficulty in solve the ambiguity problem. In order to improve this situation, this paper proposes a framework of entity type information fusing based on Transformer, which can generate word vector representation with entity type information for specific domain. There may be different entity categories for the same word, and the corresponding relationship categories are different at that time. Through deep self-attention, word vector representation is rich in entity type information, which benefits relationship extraction and ambiguity removal. The multi-layer transformer is used to realize the interaction between text features and generate a deep word vector representation with entity type information, thus effectively avoiding ambiguity. Experimental results show that our model outperforms existing methods and performs well in ambiguous contexts relative to other models. We highlight the importance of entity types in relation extraction.

Keywords: relation extraction; knowledge graph; transformer; entity type

---

1. Introduction

Relation extraction is a core task in natural language processing (NLP). The purpose is to extract the semantic relationships between two entities of the sentence and generate entity relationship triples, namely (subject, relation, object), or (s, r, o), for example (Da Vinci, born in, Italy). The acquired triples and their association relationships can provide content support for tasks such as knowledge graphs [1], question answering systems [2], recommendation systems [3], and intelligent search engines.

Early works of relational triple extraction used a pipeline method [4,5]. It first identifies the entities presented in the sentence and then confirms the existence of relationships from each entity pair. The traditional method of relationship extraction is to train a classifier based on existing data and use the trained classifier to predict the relationship. It requires a large amount of manually annotated corpus, which consumes a lot of time and energy. Based on this, [6] proposed the idea of remote supervision, extracting text features from the remotely supervised labeled data and training the relationship classification model, which effectively solved the problem of the scale of labeled data for relationship extraction. The idea of remote supervision assumes that an entity corresponds to only one relationship. With the rapid development of deep learning, some better deep entity relationship extraction algorithms have been produced in recent years. [7] used a convolutional neural network to extract entity relationships by capturing text information around each word, but this method is limited by the size of the convolution kernel, and it is difficult to obtain long-distance text information. [8] use Long Short Memory Network (LSTM) to obtain sentence semantic information to extract entity relationships. This algorithm alleviates the defects of the convolutional neural network to a certain extent, but it is still difficult to obtain the global text semantic information of long sentences. [9] proposed an entity relationship extraction model based on a dependency relationship tree, and obtained a more reliable context representation based on the dependency relationship through a long- and short-term memory network. [10] proposed a pre-training language model BERT, which has achieved good results in many natural language processing tasks. Subsequently, [11] used the output part of the BERT model to build a relational classification model R-BERT, and first used BERT in relation

extraction tasks, and explored the combination of entities and entity positions in the pre-training model. Many entities have multiple relationships between them. Therefore, [12] proposed the use of multi-instance multi-label method to model the relationship extraction to describe the situation that an entity pair that may have multiple relationships. Some works proposed joint learning methods which are joint extracting of entities and relations [13,14] which can avoid error propagation problem that entity extraction errors affect the performance of relational extraction. In general, whether focusing on the pipeline method and joint learning method of relation extraction, the existing works have improved important insights for automatically solving geometry problems.

Relation extraction needs to map contextual and entity information into machine understandable forms, and followed by the process of fusing and classification, finally presents the relation of two entities. Hence, it cannot be simply classified by contextual information or end-to-end classification methods, relationship extraction needs to be supplemented with more additional information in the model to improve precision or disambiguate. Most existing methods suffer from the same problem that the entity type information cannot be effectively used to further contribute to the performance of relationship extraction, even if the entity type is known before the relationship extraction. In ambiguous contexts, where an entity may have several different types and different relationships exist under different types. In this case, entity type information is important to disambiguate and predict the correct relationship. There are a large number of ambiguities in the NYT [15] datasets. For example, there are two types of relationships difficult to distinguish which are "contains" and "administrative_divisions". For two entities "Victoria" and "Australia", (Australia, contains, Victoria) and (Australia, administrative_divisions, Victoria) both tuples are correct. If "Australia" entity type is country then the relationship prefers "administrative_divisions", but if "Australia" entity type is location then the relationship prefers "contains". Therefore the additional entity type information can help with relationship classification. Most existing approaches cannot efficiently handle entity type information exists which can being utilized to further improve performance and disambiguation in the dataset. Additional entity information that cannot be effectively integrated into the model is the cause of this problem. Transformer [16] is widely used for feature extraction and feature fusion in natural language processing and computer vision. The self-attention and multi-head attention mechanism of Transformer can effectively integrate semantic information and additional information by the similarity between features. This gives the possibility to fuse semantic information and entity type information for relation extraction.

In this work, we propose a framework of entity type fusing for relation extraction (ETFRE) which utilizes entity type information to guide extraction. Because in some special application fields such as Science, Water Conservancy, and so on, entity type is important for relation extraction. It can also further eliminate the ambiguity in relation extraction. We break the current situation that relationship extraction cannot combine entity types by Transformer based entity type fusing mechanism. To sum up, the main contributions in this work are listed as following:

1. This paper applies entity type information to guide relation extraction which can further improve the performance of relationship classification and avoid the ambiguity in the sentence.
2. An entity type fusing framework is proposed, which can utilize entity type information and construct a deep representation of word vectors carrying entity type information through self-attention mechanism.
3. A new dataset construction format is proposed which can store entity type information in our experiments, and this approach can effectively improve the performance of relationship extraction. We conduct experiments on three public datasets, and the weighted F1 value of our model achieves an absolute improvement by +1.5%, +1.0%, +1.2% on CL, NYT, and SciERC dataset respectively.

2. Related Work

As an important step for constructing large scale knowledge graph such as Freebase [17], relational triple extraction has been extensively studied over the decades. Existing researches on relational

triple extraction can be mainly divided into two categories: pipeline method and joint learning methods.

Pipeline Method This approach uses two steps to generate relational triple. Firstly, it identifies the entities presenting in the sentence by named entity recognition (NER) [18] and secondly confirms the existence of relationships from each entity pair by relation classification (RC) on pairs of extracted entities. In this case, two entities have been marked out and the relationship extraction is considered as a classification problem. [15] introduce multi-instance learning (MIL) which claimed that if a relationship exists between two entities, at least one sentence that contains the two entities may express the corresponding relation. [19] made further improvements by introducing an attention mechanism that attends to relevant information in a bag of sentences. This sentence-level attention mechanism for MIL inspired numerous subsequent works [20]. Wu et al. used the BERT model for embedding words into features based on this building a relational classification model R-BERT which first used BERT in relation extraction tasks, and explored the combination of entities and entity positions in a single model. They usually only pass the entity location information to the relationship extraction task, and information like entity type is not used effectively.

Joint Learning Methods Many joint models that aim to extract entities and relations jointly used to ease error propagation problem. Zheng et al. [13] propose a joint relational triple extraction framework based on potential relation and global correspondence by three sub-tasks which are relation judgement, entity extraction and subject-object alignment. Although the joint approach allows a single model to extract triples, it also still does not take advantage of the additional entity type information to bring performance gains to relationship extraction. [21] propose a joint training model that combines a knowledge graph with an attention mechanism and MIL. Experiments on relation extraction and entity link prediction show that models trained under their joint framework are significantly improved in comparison with other baselines. Wei et al. [22] proposed a cascade binary tagging framework which model the relations as functions that map subjects to objects.

Early relation extraction methods mainly used to find the law of the text and formulate a series of rules to extract relations, such as the method based on rules and dictionaries. This type of extraction method has a high evaluation index, but it needs manual construction, and the cost is high, and the text size that can be processed is also small. Because of these limitations, supervised relation extraction methods have attracted the attention of researchers with machine learning methods, such as eigenvector-based method and kernel functions. [23] use naive bayesian algorithm and perceptron algorithm constructed a feature-based Chinese term relation extraction model. Rink [24] construct relational classifiers based on support vector machine(SVM) by combining various linguistic resources. However, the performance of traditional machine learning models is very dependent on the scale and quantity of manually labeled feature data, so a method that can automatically extract features is needed. Deep learning has the characteristics of self-learning, which can automatically extract features, reduce dependence on manual work, and extract large-scale text data. Deep learning methods mainly include supervision and remote supervision. [25] propose a model that uses BioBERT [26], a pretrained transformer based on bert, for sentence encoding. The author proves that the pre training of biomedical corpus for Bert is the key to its application in biomedical field. They leverage MIL with entity-marking methods following R-BERT and achieve the best performance when the directionality of extracted triples is matched to the directionality from the UMLS knowledge graph. The methods based on deep learning have greatly promoted the development of the field of relation extraction.

Our framework use pipeline method to extract relationship based on a BERT embedder like R-BERT and enables entity type information to be used effectively for supervisory relationship classification. Moreover, we fuse entity type information into features, which makes it crucially different from previous works.

3. Problem Statements

Relation extraction is a key task for knowledge graph construction and natural language processing, which aims to extract meaningful relational information between entities from plain texts. This task can be divided into three modules for pipeline method: Entity recognition and Relation trigger word identification; and finally Relation extraction module.

Entity recognition refers to the recognition of entities with specific meaning in the text, mainly including person names, place names, proper nouns, etc.; Relation trigger word identification, refers to the classification of the words that trigger the entity relationship, and recognizes whether it is a trigger word or a non-trigger Words determine whether the extracted relationship is positive or negative; Relation extraction refers to extracting semantic relationships between entities, such as employees, birthplaces, products, etc., from the identified entities. Taking the sentence "Lu Xun was born in Shaoxing" as an example, the sentence is first preprocessed to identify the named entities "Lu Xun" and "Shaoxing", and then "was born in" as the relationship trigger word indicates that there may be some kind of relationship between these two entities. Finally, through the judgment of the relation extraction model, it is concluded that there is a "Birthplace" relationship between the two entities.

Relation extraction is a text classification problem. Compared with other tasks such as sentiment classification and news classification. Relation extraction has three main characteristics: 1) it involves many fields, and the construction of relational models is complex; 2) data sources are extensive, involving structured, semi-structured, and non-structured; 3) the types of relationships are diverse and complex, and noisy data is unavoidable. However, the existing relationship extraction methods in various fields do not use the existing entity category information, making effective information unavailable. In response to this problem, this paper proposes a transformer-based entity type information fusion framework to improve relation extraction performance.

4. Method

This section presents the proposed model of entity type fusing for relation extraction. As shown in Figure 1, relation extraction is a three-step process: first, we use pre-trained BERT to embed each word into a full of semantic information vector; second, for each vector, we concatenate entity type information with each word vector to further get the representation which is helpful for relationship extraction. A single layer Transformer is proposed to utilize self-attention to dig deeper expressions of words that carry entity information; third, the multi-layer perception has been used to condense the word vectors of a sentence into a column vector which will feed into a classifier layer to produce the probability distribution of relationship categories. Corresponding to the three processing steps, there are three key parts which are BERT Embedder, Transformer Fusing and Relation Classifier. At the down side of model, BERT Embedder has been used to embed each word and '[CLS]' into vectors. Before feed sentence into BERT Embedder, '[CLS]' has been insert to the beginning of sentence and special tokens have been insert to beginning and end of each entity which mark location of entity. In the middle of the model diagram, Transformer Fusing has been used to construct a deep representation of word vectors which can carrying entity type information. At this part, each output vector of BERT except '[CLS]' token's embedding has been concatenate with another vectors which are all zero vector or entity type embedding. For each entity, each vector of entity embedding from BERT has been concatenate with corresponding entity type embedding. For other words, they simply concatenate with all zero vector. Then those concatenated vectors are fed into a single layer transformer to get deeper representation of word with entity type information. At the top of the model diagram, Relation Classifier have been used to produce probability distribution of prediction relation. In this module, a multi-layer perception have been used to fuse output from Transformer into a single vector. This single vector has been concatenate with embedding of '[CLS]' and finally feed into a fully connected layer and a softmax layer to produce relational probability.
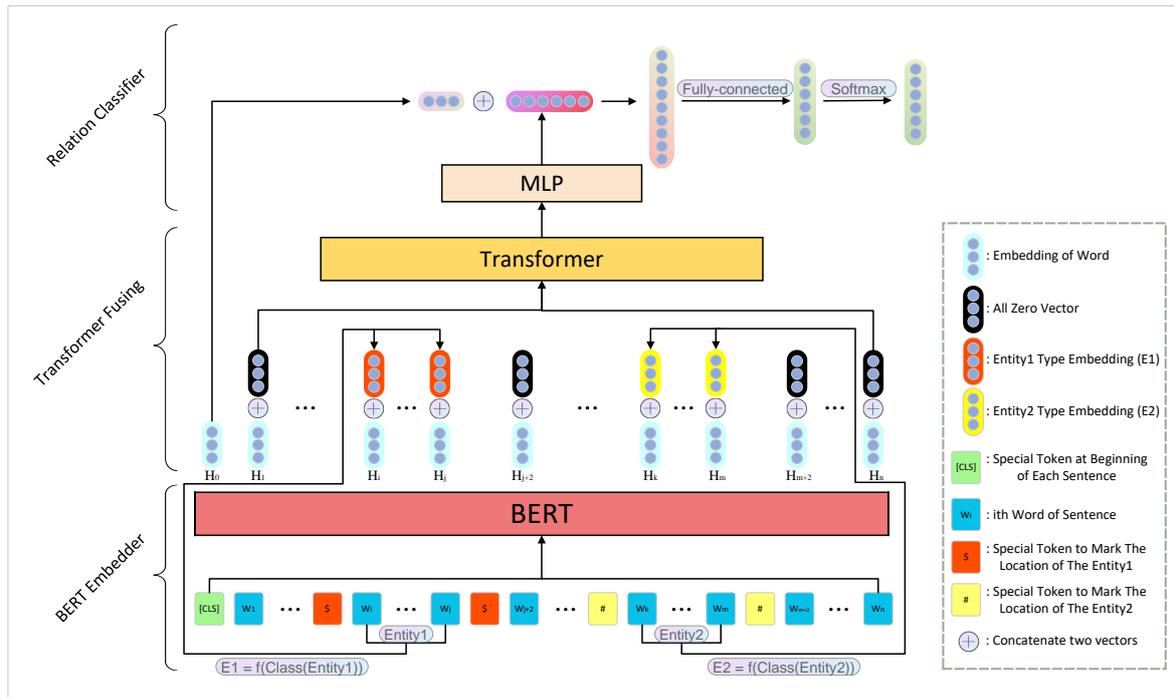
Figure 1. This is a wide figure.

4.1. BERT Embedder

For the next stage transformer fusing, two special token '$' and '#' have been appended into the given sentence s with two relevant entities e_1 and e_2. At the beginning and end of first entity '$' have been inserted, and at both the beginning and end of second entity '#' have been appended. We also add '[CLS]' to the beginning of every sentience which will input into BERT and return a single vector which treats as encoding of the sentence. For example, after insertion of the special tokens, for a sentence with target entities "Girls" and "lovely children" will become to:

"[CLS] Fortunately, both boys and girls are mature. It's not 20 years ago. Boys see through nature. Now they look down on this life. $Girls$ have a good life. They have #lovely children# and loving husbands."

Given a sentence s with entity e_1 and e_2 which input into BERT, we treat final hidden state output of BERT module is H which include embedding of '[CLS]' and every word of sentience.

4.2. Transformer Fusing

Each entity has its class. Given one-hot encoding of entity class, we embed each class c^i into a vector space by $c_i = Mc^i$, where M is an embedding matrix. Finally each entity class is embed into $C = [\mathbf{c}_0; \ldots; \mathbf{c}_n]$, which n is the number of overall entity class. When we get the output of BERT embedder, each word of vector dimension is 768 which will expand the dimension by 128. The expanded value of each word embedding is all zero except for entity embedding. The position of the entity is pre-marked by '$' and '#'. We expand each entity class embedding into entity word embedding. For example, there are two words in e_2 which are 'lovely' and 'children', the word embedding of these two words by BERT is h_i and h_{i+1}. Entity type of e_2 is 'Person-Nominal' and it's type embedding is c_j. Then the expanded embedding of 'lovely' is $[c_j, h_i]$ and 'children' is $[c_j, h_{i+1}]$. To get deeper information expression of each word with entity class embedding. A single layer Transformer [16] have been posed to fuse each word expanded embedding. By self-attention, a deeper expression of each word with its class information have been generated. Finally, each word with it's entity class information have been encode into $T = [\mathbf{t}_1; \ldots; \mathbf{t}_n]$ which n is the length of sentence. It is important to note that special marks '[CLS]' are not input into the Transformer.

4.3. Relation Classifier

When we get T from Transformer fusing, first a multi-layer perception (MLP) have been used to fuse T into a single vector. We also use the activation operation and a fully connected layer to makes the fused vectors more compatible with the next operations, which is formally expressed as:

$$t' = W_0(\tanh(mlp(T))) + b_0 \tag{1}$$

Embedding of '[CLS]' is also processed by an activation operation and a fully connected layer then we get s'. We concatenate s' and t' then add a fully connected layer which can be expressed as:

$$h' = W_1[concat(s', t')] + b_1 \tag{2}$$

Finally, in order to obtain the relational probability distribution, a softmax layer is added to the end, which is expressed as below:

$$p = softmax(h') \tag{3}$$

Where p is probability distribution of predicted relationship between e_1 and e_2. Matrices W_0, W_1 are trainable parameters and b_0, b_1 are bias vectors.

5. Experiments

5.1. Experimental Setup

We evaluate the proposed approach on the open dataset of Chinese Literature (CL) [27], NYT [28] and SciERC [29]. The dataset of CL is constructed by Brat [30] which provide an intuitive and fast way to create text-bound and relational annotations. The SciERC dataset is created for scientific information extraction and collected from 500 AI paper abstracts. For fair and comprehensive comparison, we also follow [13] to evaluate our model on public datasets NYT. The relation types and entity types of these three datasets are shown in the Table 1.

Table 1. Statistics of the datasets.

Dataset	Relation types	Entity types
CL	10	15
NYT	24	14
SciERC	7	6

NYT and SciERC datasets are proposed in a json format. We first transform those datasets into format of tab-separated values (TSV) for model input. As Figure 2 shows, each data in dataset have four parts which is processed sentence, relation type, and two entity types. For a sentence s with two target entities e_1 and e_2, we capture the location information of the two entities at both the beginning and end of the first entity. We insert special tokens which are '<e1>' and '</e1>', and at both the beginning and end of the second entity. Then, we insert special tokens which are '<e2>' and '</e2>'. Relation between Bed and Villa is 'Located' and entity type for Bed is 'Ting-Nominal' and entity type for Villa is 'Location-Nominal'.

Relation Type	First Entity Type	Second Entity Type
Located(e2,e1)	Location-Nominal	Thing-Nominal

In 1883, Turgenev lay on a <e2>Bed</e2> in the <e1>Villa</e1> to bid farewell to the world, ending his 40 year legendary relationship with Polina.

Figure 2. An example data format which we transformed from Chinese Literture.

After datasets processed, they are randomly partitioned into a training set, a validation set, and a test set, number of each set as shown in Table 2.

Table 2. Datasets splitting.

Dataset	Train	Test	Validation
CL	11553	2889	2889
NYT	80000	8106	8233
SciERC	3215	974	812

For fair comparison, Precision(Prec.), Recall(Rec.) and F1-score have been used to measure model performance. To account for label imbalance of each relation, we calculate metrics for each label, and find their average weighted by the number of true instances for each label.

5.2. Implementation Details

We use validation set to determined hyper-parameters. Batch size have been set as 16. Because of batch operation, sentences need to be filled with padding which max sentence length is 384. Excessively long sentences will be truncated without discarding entities. Our model is implemented with PyTorch and the network weights are optimized with Adam. Adam learning rate is set to $2e^{-5}$. Early stopping mechanism have been used to prevent model over-fitting. Training process will be stopped when the performance of validation set does not get any improvement for at least 5 consecutive epochs. The single layer Transformer Fusing of hidden size is 896. Hidden size of BERT is 768 and number of stacked bidirectional Transformer blocks N is 12. We use pre-trained BERT model which is [bert–base–chinese] for Chinese Literature dataset and [bert-base-uncased] for NYT and SciERC. Both pre-trained BERT model contains 110M parameters. We add dropout for each activation layers which dropout rate is 0.1. We use A6000 to train the model for at most 100 epochs and choose the model with the best performance on the validation set to output results on the test set.

5.3. Experimental Result

Table 3 shows the results of our approach ETFRE for relational triple extraction and compares with previous results on CL, NYT, and SciERC datasets.

Table 3. Results of different methods on Chinese Literature, NYT and SciERC datasets. **Bold** marks the highest score and ‡ marks the results reported by the original papers.

Method	CL			NYT			SciERC		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
R-BERT [11]	78.0	78.0	77.8	77.9	69.8	69.2	86.2	86.3	85.7
CasRel‡ [22]	-	-	-	89.7	89.5	89.6	-	-	-
TPLinker‡ [31]	-	-	-	91.3	92.5	91.9	-	-	-
PRGC‡ [13]	-	-	-	93.3	91.9	92.6	-	-	-
PURE [32]	-	-	-	-	-	-	**91.1**	68.4	77.8
Ours	**79.4**	**79.7**	**79.3**	**95.0**	**94.1**	**93.6**	87.2	**87.2**	**86.9**

We can see that our ETFRE model beats majority of the baseline methods in terms of three evaluation metrics. The weighted F1 value of our model improves +1.5%, +1.0%, +1.2% on CL, NYT, and SciERC respectively. This shows that our entity type fusing approach by Transformer is useful for relation extraction.

We sample some results predicted by our model and R-BERT model as shown in Appendix A.1. We selected data with ambiguity from the NYT dataset. These sentences with ambiguity are difficult to predict correctly for R-BERT, but our model can effectively make predictions. This highlights the

importance of entity types in relationship prediction. It also verifies that our model can effectively fuse entity type information for relationship prediction.

5.4. Ablation Study

In this section, We further want to understand the specific contributions of the pre-trained BERT component. For this purpose, We use random initialization of BERT for comparison. To measure how powerful BERT is in the word embedding, we replaced BERT embedder by LSTM. The word representation of Transformer fusing results are visualized by t-SNE to show some relevances. The ablation experiments to demonstrate the effectiveness of each component as shown in Table 4. LSTM-Replace is our model replacing embedder with LSTM. BERT-Random is the framework where all parameters of BERT are randomly initialized and Type-strip is striping entity type in transformer fusing which only fuse BERT embedder outputs.

Table 4. Ablation experiment on NYT and Chinese Literature dataset.

Dataset	Model	Prec.	Rec.	F1
NYT	Ours	95.0	**94.1**	93.6
	LSTM-Replace	88.8	86.6	86.3
	BERT-Random	**95.1**	94.0	**93.7**
	Type-Strip	90.8	91.0	90.9
CL	Ours	**79.4**	**79.7**	**79.3**
	LSTM-Replace	73.1	73.0	72.8
	BERT-Random	79.0	79.0	78.8
	Type-Strip	76.7	77.1	78.3

The results indicate pre-trained BERT is rich in semantic information for relation extraction. The LSTM embedder drop a large gap of preformances compare with ETFRE which indicate the importance of BERT in the word embedding.

To explore the effect of entity type on word vectors, we randomly choose some word representation by Transformer fusing in training set of SciERC to visualize. Firstly, principal component analysis (PCA) is used to reduce dimensions to a reasonable amount (200) then we use t-SNE to reduce dimensions of representation into two dimensions for visualizations as shown in Figure 3. The distance between any two word points illustrates the similarity of the syntax, semantics, and entity types. For different entity type, we use different color to better distinguish. For example, the blue point, e.g., "accuracy", "relations", "names" and, "shape" are the same type of *"OtherScientificTerm"*. It is obvious from the figure that after adding the entity type information to the word vector by Transformer fusing, the words are presented in clusters with same entity types for majority. The distance of entities of the same type are more close to each other, and illustrate the more similar their word vector representations are on the graph. This phenomenon indicates Transformer fusing with entity type is valid for word representation. The word representations contain entity information can effectively facilitate relationship classification.

To demonstrate the importance of entity types for relationships, we use heat map to display number of relations at different entity type pairs. As Figure 4 shows, the horizontal axis indicates the different entity type pairs. Entity names are simplified for ease of diagramming which 'L' is short of "Location", 'P' is short of "Person" and 'T' is short of "Thing". The vertical axis represents the different types of relations. Each value in the heat map is number of corresponding relations at different entity type pairs of dataset. For the relations and entity type pairs with less than 100 occurrences in the Chinese Literature dataset were omitted and not shown in the heat map. For each relation types, their entity type pair mainly focused on two to three type pairs. Each type pair corresponds to a more focused relationship, so with the entity type pair information, it can effectively narrow down the

decision space of relation classification. In other words, narrowing down the possible decision space from all the relationship categories present in the dataset to two to three.
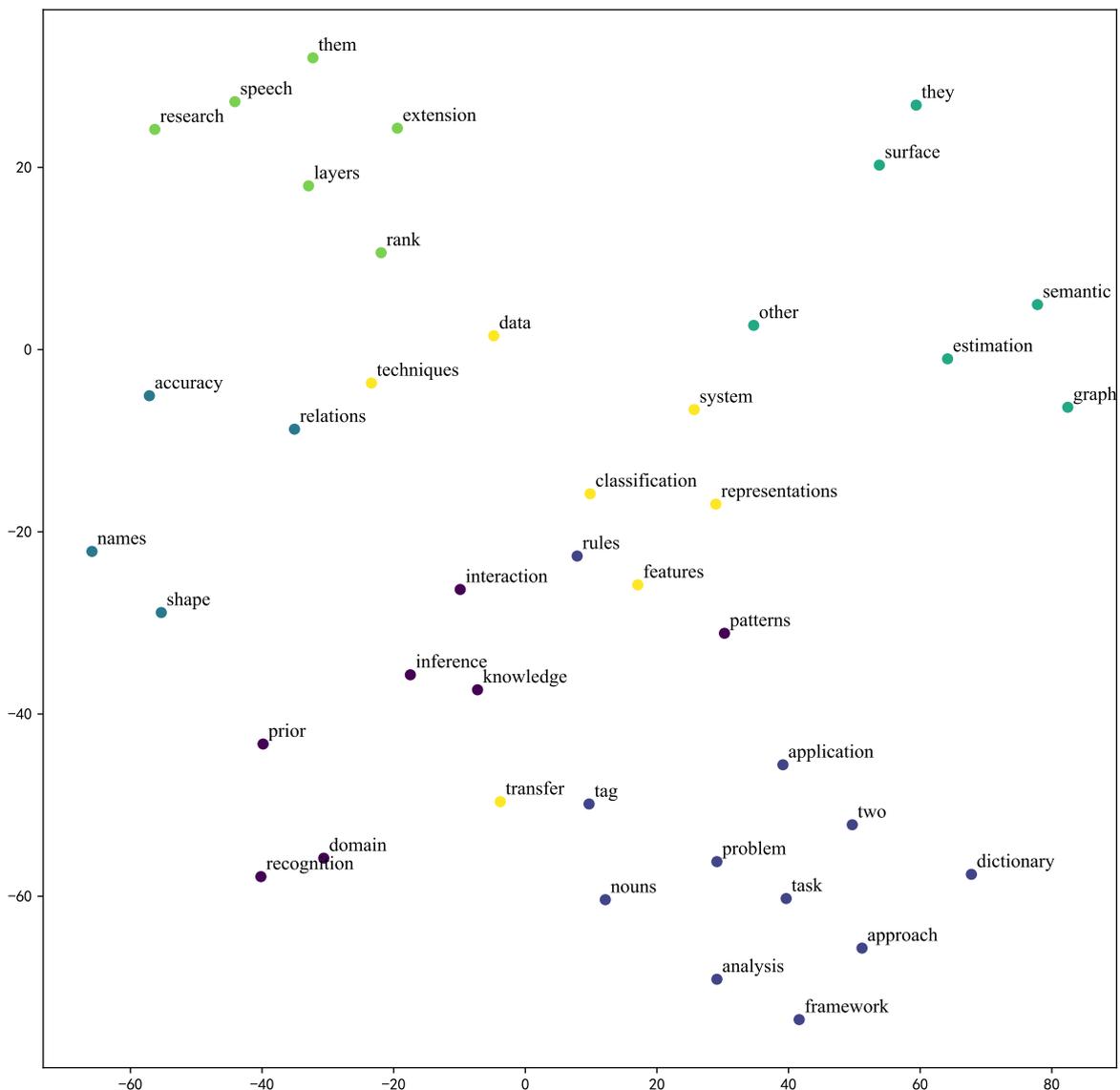

Figure 3. Visualizations of the learned word representations by t-SNE.

For NYT dataset, we also draw heat map to display distributions as Figure 5 shows and omit relations less than 400 occurrences and entity type pairs less than 100 occurrences which are not shown in the figure. For each relation in the NYT, they mainly focus on two entity type pairs. This concentration trend is more pronounced than Chinese Literature dataset. This highlights that the current popular relational classification datasets present this common phenomenon that each relationship is usually focused on only a small number of entity class pairs. We think that the presence of this quality allows additional entity type information to make the multi-classification search space smaller and make a smaller variety of candidate relations, which could be the reason for the better performance of our model.
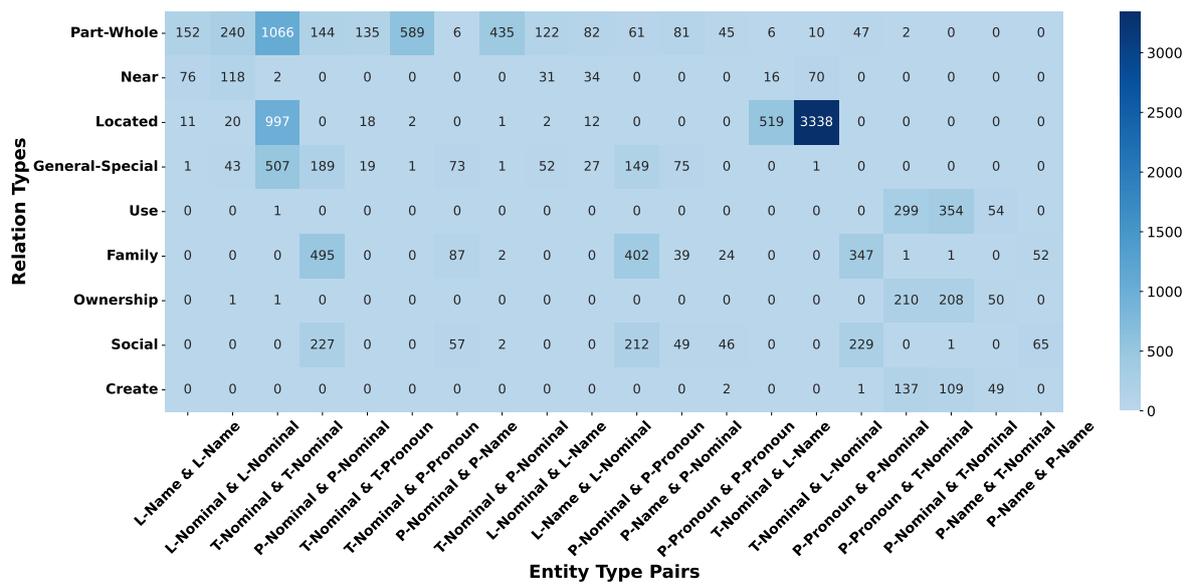
Figure 4. Heat map of the each relation in different entity type pairs distribution in Chinese Literature.
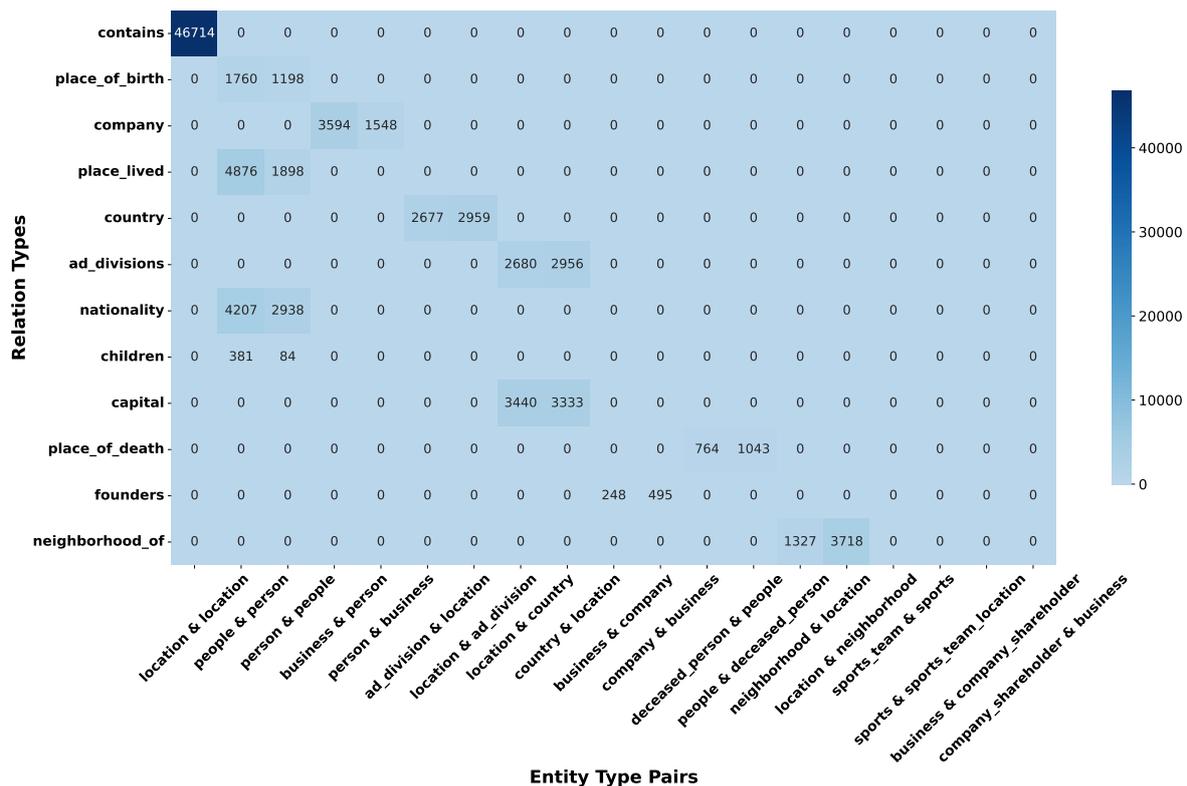

Figure 5. Heat map of the each relation in different entity type pairs distribution in NYT.

6. Conclusion

In this paper, we propose an end-to-end model for relation extraction which utilizes entity type information. A transformer based fusing mechanism has been posed to excavate deeper expression of word features with entity type information. The experiment results on the CL, NYT, and SciERC datasets demonstrated the effectiveness of our framework with the weighted F1 score improved by +1.5%, +1.0%, +1.2% respectively. We filled the gap in which the existing model cannot use entity category information to further improve performance when almost methods focus on overlapping

triple problems. We also demonstrate that entity type information performs better in the ambiguous environment of relation extraction. Ablation study demonstrated the importance of the BERT embedder. In addition, we showed that Transformer fusing can effectively fuse entity types into word representation which vectors of the same entity type words will be close together in the semantic space. Statistical information on existing datasets highlights the importance of entity type information in relation extraction.

In the future, we would like to generalize the entity type fusing idea and explore its performance on other open access datasets. We will explore how the richness and completeness of entity types and the relationships in the datasets and build a more general framework for relationship extraction.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Enquiries about data availability should be directed to the authors.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A.

Appendix A.1. Examples of Ambiguous Sentences

Somewhat chastened by his retreat in the polls , Mr. Blair acknowledged that Britons had turned against him in part over accusations that he led them into a war in <e1>Iraq</e1> on dubious legal grounds and on the false premise that Saddam <e2>Hussein</e2> presented a direct threat because of a supposed arsenal of unconventional weapons that was never found .
e_1 **type: people;** e_2 **type: deceased_person**
✓**Our Model: place_of_death(e_2,e_1)**
✗**R-BERT Model: place_of_birth(e_2,e_1)**

Somewhat chastened by his retreat in the polls , Mr. Blair acknowledged that Britons had turned against him in part over accusations that he led them into a war in <e1>Iraq</e1> on dubious legal grounds and on the false premise that Saddam <e2>Hussein</e2> presented a direct threat because of a supposed arsenal of unconventional weapons that was never found .
e_1 **type: people;** e_2 **type: person**
✓**Our Model: place_of_birth(e_2,e_1)**
✓**R-BERT Model: place_of_birth(e_2,e_1)**

Kerry <e1>Packer</e1>, who became Australia 's richest man by turning a magazine and television inheritance worth millions into a diverse business worth billions , died yesterday in <e2>Sydney</e2>.
e_1 **type: people;** e_2 **type: deceased_person**
✓**Our Model: place_of_death(e_1,e_2)**
✗**R-BERT Model: place_lived(e_1,e_2)**

Kerry <e1>Packer</e1>, who became Australia 's richest man by turning a magazine and television inheritance worth millions into a diverse business worth billions , died yesterday in <e2>Sydney</e2>.
e_1 **type: people;** e_2 **type: person**
✗**Our Model: place_of_birth(e_1,e_2)**
✓**R-BERT Model: place_lived(e_1,e_2)**

With an eye to enlivening some furniture in my home office , I have been stocking up on 1960 's and 70 's fabrics from Retro Age Vintage Fabric , a dealer in <e1>Victoria</e1>, <e2>Australia</e2>, that sells everything from trippy Art Nouveau-style patterns to Lilly Pulitzer-style florals .

e_1 **type: location;** e_2 **type: country**

✓ **Our Model: administrative_divisions(e_2,e_1)**

✓ **R-BERT Model: administrative_divisions(e_2,e_1)**

With an eye to enlivening some furniture in my home office , I have been stocking up on 1960 's and 70 's fabrics from Retro Age Vintage Fabric , a dealer in <e1>Victoria</e1>, <e2>Australia</e2>, that sells everything from trippy Art Nouveau-style patterns to Lilly Pulitzer-style florals .

e_1 **type: location;** e_2 **type: location**

✓ **Our Model: contains(e_2,e_1)**

✗ **R-BERT Model: administrative_divisions(e_2,e_1)**

The husband , Joseph Vione , 43 , who had been living in Garden City , says in papers filed last week in State Supreme Court in <e1><e2>Manhattan</e2></e1> that the Rev . Thomas K. Tewell , 56 , the senior pastor of the Fifth Avenue Presbyterian Church in Midtown Manhattan , used confidential information obtained during marriage counseling to seduce Mr. Vione 's wife , Rachel , 42 .

e_1 **type: location;** e_2 **type: location**

✓ **Our Model: contains(e_1,e_2)**

✗ **R-BERT Model: neighborhood_of(e_1,e_2)**

The husband , Joseph Vione , 43 , who had been living in Garden City , says in papers filed last week in State Supreme Court in <e1><e2>Manhattan</e2></e1> that the Rev . Thomas K. Tewell , 56 , the senior pastor of the Fifth Avenue Presbyterian Church in Midtown Manhattan , used confidential information obtained during marriage counseling to seduce Mr. Vione 's wife , Rachel , 42 .

e_1 **type: location;** e_2 **type: neighborhood**

✓ **Our Model: neighborhood_of(e_1,e_2)**

✓ **R-BERT Model: neighborhood_of(e_1,e_2)**

New <e1>Zealand</e1> : <e2>Marlborough</e2>, Hawke 's Bay and Central Otago Despite having vines dating to 1819 , New Zealand wines were not known globally until the mid-1980 's , when their vibrant , minerally , tropical-fruit-filled sauvignon blanc set the wine world abuzz .

e_1 **type: location;** e_2 **type: country**

✓ **Our Model: administrative_divisions(e_1,e_2)**

✗ **R-BERT Model: country(e_2,e_1)**

New <e1>Zealand</e1> : <e2>Marlborough</e2>, Hawke 's Bay and Central Otago Despite having vines dating to 1819 , New Zealand wines were not known globally until the mid-1980 's , when their vibrant , minerally , tropical-fruit-filled sauvignon blanc set the wine world abuzz .

e_1 **type: location;** e_2 **type: location**

✓ **Our Model: contains(e_1,e_2)**

✗ **R-BERT Model: country(e_2,e_1)**

But from downtown 's Little Tokyo , home to many izakaya hole-in-the-walls , to West Los <e1><e2> Angeles</e2></e1>, where a new-wave izakaya serves duck breast marinated in sake along with Basque sheep 's milk cheese , Los Angeles may have the most inventive permutations of izakaya-style restaurants in the United States .

e_1 **type: location;** e_2 **type: location**

✓ **Our Model: contains(e_1,e_2)**

✗ **R-BERT Model: neighborhood_of(e_1,e_2)**

But from downtown 's Little Tokyo , home to many izakaya hole-in-the-walls , to West Los <e1><e2> Angeles</e2></e1>, where a new-wave izakaya serves duck breast marinated in sake along with Basque sheep 's milk cheese , Los Angeles may have the most inventive permutations of izakaya-style restaurants in the United States .

e_1 **type: location;** e_2 **type: neighborhood**

✓ **Our Model: neighborhood_of(e_1,e_2)**

✓ **R-BERT Model: neighborhood_of(e_1,e_2)**

1 <e1>Bali</e1> Suspects Elude Capture Three men wanted in connection with bombings in Bali that killed 222 people over three years , narrowly escaped capture in the last two days , two in the Philippines and one in <e2>Indonesia</e2>, according to officials in those countries .

e_1 type: location; e_2 type: location

✓ **Our Model: contains(e_2,e_1)**

✗ **R-BERT Model: administrative_divisions(e_2,e_1)**

1 <e1>Bali</e1> Suspects Elude Capture Three men wanted in connection with bombings in Bali that killed 222 people over three years , narrowly escaped capture in the last two days , two in the Philippines and one in <e2>Indonesia</e2>, according to officials in those countries .

e_1 type: location; e_2 type: country

✓ **Our Model: administrative_divisions(e_2,e_1)**

✓ **R-BERT Model: administrative_divisions(e_2,e_1)**

An exhibition of Mr. Parr 's images opened at Danziger Projects in Chelsea this week , showing parallels between staged photography of models wearing designer collections and candid ones of people he encountered in <e1>Dakar</e1>, <e2>Senegal</e2> and Cuba .

e_1 type: location; e_2 type: location

✓ **Our Model: contains(e_2,e_1)**

✓ **R-BERT Model: capital(e_2,e_1)**

An exhibition of Mr. Parr 's images opened at Danziger Projects in Chelsea this week , showing parallels between staged photography of models wearing designer collections and candid ones of people he encountered in <e1>Dakar</e1>, <e2>Senegal</e2> and Cuba .

e_1 type: location; e_2 type: country

✓ **Our Model: capital(e_2,e_1)**

✗ **R-BERT Model: capital(e_2,e_1)**

By contrast , cities in the export-oriented Guangdong <e1>Province</e1> in southeastern <e2>China</e2> raised monthly minimum wages this summer by 18 percent , to 70to100 a month , after factories reported that they had one million more jobs than workers to fill them .

e_1 type: location; e_2 type: location

✓ **Our Model: contains(e_2,e_1)**

✗ **R-BERT Model: administrative_divisions(e_2,e_1)**

By contrast , cities in the export-oriented Guangdong <e1>Province</e1> in southeastern <e2>China</e2> raised monthly minimum wages this summer by 18 percent , to 70to100 a month , after factories reported that they had one million more jobs than workers to fill them .

e_1 type: location; e_2 type: country

✓ **Our Model: administrative_divisions(e_2,e_1)**

✓ **R-BERT Model: administrative_divisions(e_2,e_1)**

Leading the parade is actually the Duke of <e1>Saxony</e1> -LRB- below -RRB- , whose suit and that of his mount were made by one of <e2>Germany</e2>'s leading armorers , Kunz Lochner , in 1548 .

e_1 type: location; e_2 type: location

✓ **Our Model: contains(e_2,e_1)**

✗ **R-BERT Model: administrative_divisions(e_2,e_1)**

Leading the parade is actually the Duke of <e1>Saxony</e1> -LRB- below -RRB- , whose suit and that of his mount were made by one of <e2>Germany</e2>'s leading armorers , Kunz Lochner , in 1548 .

e_1 type: location; e_2 type: country

✓ **Our Model: administrative_divisions(e_2,e_1)**

✓ **R-BERT Model: administrative_divisions(e_2,e_1)**

<e1>Canada</e1> can thank British <e2>Columbia</e2> for much of its talent .

e_1 type: location; e_2 type: country

✓ **Our Model: administrative_divisions(e_1,e_2)**

✗ **R-BERT Model: country(e_2,e_1)**

<e1>Canada</e1> can thank British <e2>Columbia</e2> for much of its talent .

e_1 **type: location;** e_2 **type: location**

✓**Our Model: contains(e_1,e_2)**

✗**R-BERT Model: country(e_2,e_1)**

Even Larry <e1>Page</e1>, the <e2>Google</e2> co-founder who enjoyed rock-star treatment at the World Economic Forum in Davos , Switzerland , seemed to have reverted to regular-guy status in Sun Valley – as regular as you can be with billions of dollars of Google stock .

e_1 **type: business;** e_2 **type: company_shareholder**

✓**Our Model: major_shareholder_of(e_1,e_2)**

✗**R-BERT Model: company(e_1,e_2)**

Even Larry <e1>Page</e1>, the <e2>Google</e2> co-founder who enjoyed rock-star treatment at the World Economic Forum in Davos , Switzerland , seemed to have reverted to regular-guy status in Sun Valley – as regular as you can be with billions of dollars of Google stock .

e_1 **type: business;** e_2 **type: person**

✓**Our Model: company(e_1,e_2)**

✓**R-BERT Model: company(e_1,e_2)**

References

1. Pujara, J.; Miao, H.; Getoor, L.; Cohen, W. Knowledge Graph Identification. The Semantic Web – ISWC 2013; Alani, H.; Kagal, L.; Fokoue, A.; Groth, P.; Biemann, C.; Parreira, J.X.; Aroyo, L.; Noy, N.; Welty, C.; Janowicz, K., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; pp. 542–557.

2. Dong, L.; Wei, F.; Zhou, M.; Xu, K. Question Answering over Freebase with Multi-Column Convolutional Neural Networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Association for Computational Linguistics: Beijing, China, 2015; pp. 260–269. doi:10.3115/v1/P15-1026.

3. Zhang, F.; Yuan, N.J.; Lian, D.; Xie, X.; Ma, W.Y. Collaborative Knowledge Base Embedding for Recommender Systems. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: New York, NY, USA, 2016; p. 353–362. doi:10.1145/2939672.2939673.

4. Zelenko, D.; Aone, C.; Richardella, A. Kernel Methods for Relation Extraction. *Journal of Machine Learning Research* **2003**, *3*, 1083–1106.

5. Chan, Y.S.; Roth, D. Exploiting syntactico-semantic structures for relation extraction. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 551–560.

6. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant supervision for relation extraction without labeled data. 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; Association for Computational Linguistics: Suntec, Singapore, 2009; pp. 1003–1011.

7. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation Classification via Convolutional Deep Neural Network. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers; Dublin City University and Association for Computational Linguistics: Dublin, Ireland, 2014; pp. 2335–2344.

8. Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; Jin, Z. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 1785–1794. doi:10.18653/v1/D15-1206.

9. Miwa, M.; Bansal, M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Berlin, Germany, 2016; pp. 1105–1116. doi:10.18653/v1/P16-1105.

10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4171–4186. doi:10.18653/v1/N19-1423.

11. Wu, S.; He, Y. Enriching pre-trained language model with entity information for relation classification. Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 2361–2364.

12. Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; Weld, D.S. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; Association for Computational Linguistics: Portland, Oregon, USA, 2011; pp. 541–550.

13. Zheng, H.; Wen, R.; Chen, X.; Yang, Y.; Zhang, Y.; Zhang, Z.; Zhang, N.; Qin, B.; Xu, M.; Zheng, Y. PRGC: Potential relation and global correspondence based joint relational triple extraction. *arXiv preprint arXiv:2106.09895* **2021**.

14. Fu, T.J.; Ma, W.Y. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction. ACL, 2019.

15. Riedel, S.; Yao, L.; Mccallum, A.K. Modeling relations and their mentions without labeled text. Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III, 2010.

16. Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; Zhong, J. Attention is all you need in speech separation. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 21–25.

17. Bollacker, K. Freebase : A collaboratively created graph database for structuring human knowledge. *Proc. SIGMOD' 08* **2008**.

18. Ratinov, L.; Roth, D. Design challenges and misconceptions in named entity recognition. Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009), 2009, pp. 147–155.

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

20. Luo, B.; Feng, Y.; Wang, Z.; Zhu, Z.; Huang, S.; Yan, R.; Zhao, D. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. *arXiv preprint arXiv:1705.03995* **2017**.

21. Han, X.; Liu, Z.; Sun, M. Neural knowledge acquisition via mutual attention between knowledge graph and text. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, Vol. 32.

22. Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; Chang, Y. A novel cascade binary tagging framework for relational triple extraction. *arXiv preprint arXiv:1909.03227* **2019**.

23. Xia, S.X.S.; Lehong, D.L.D. Feature-Based Approach to Chinese Term Relation Extraction. International Conference on Signal Processing Systems, 2009.

24. Rink, B.; Harabagiu, S. UTD: Classifying semantic relations by combining lexical and semantic resources. *Association for Computational Linguistics* **2010**.

25. Amin, S.; Dunfield, K.A.; Vechkaeva, A.; Neumann, G. A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction. *arXiv e-prints* **2020**.

26. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**.

27. Lütkebohle, I. Chinese-Literature-NER-RE-Dataset. https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset/tree/master/relation_extraction, 2008. [Online; accessed 19-July-2008].

28. Riedel, S.; Yao, L.; McCallum, A. Modeling Relations and Their Mentions without Labeled Text BT - Machine Learning and Knowledge Discovery in Databases. *Ecml* **2010**, pp. 148–163.

29. Luan, Y.; He, L.; Ostendorf, M.; Hajishirzi, H. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602* **2018**.

30. Stenetorp, P.; Pyysalo, S.; Topić, G.; Ohta, T.; Ananiadou, S.; Tsujii, J. brat: a Web-based Tool for NLP-Assisted Text Annotation. Proceedings of the Demonstrations Session at EACL 2012; Association for Computational Linguistics: Avignon, France, 2012.

31.   Wang, Y.; Yu, B.; Zhang, Y.; Liu, T.; Zhu, H.; Sun, L. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415* **2020**.

32.   Zhong, Z.; Chen, D. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812* **2020**.