Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200755

A Brief Review of Relation Extraction Based on Pre-Trained Language Models

Tiange Xu and Fu Zhang[1]
School of Computer Science & Engineering, Northeastern University, Shenyang, China

Abstract. Relation extraction is to extract the semantic relation between entity pairs in text, and it is a key point in building Knowledge Graphs and information extraction. The rapid development of deep learning in recent years has resulted in rich research results in relation extraction tasks. At present, the accuracy of relation extraction tasks based on pre-trained language models such as BERT exceeds the methods based on Convolutional or Recurrent Neural Networks. This review mainly summarizes the research progress of pre-trained language models such as BERT in supervised learning and distant supervision relation extraction. In addition, the directions for future research and some comparisons and analyses are discussed in our whole survey. The survey may help readers understand and catch some key techniques about the issue, and identify some future research directions.

Keywords. Relation Extraction, Pre-trained Language Models, Review

1. Introduction

Relation extraction plays an important role in the fields of Knowledge Graph, Machine Translation, Text Data Mining, and Information Extraction. Relation extraction has attracted widespread attention in international conferences on AI and NLP, and has achieved a series of research results. The traditional relation extraction methods are divided into rule-based learning, supervised learning, weakly supervised learning, and unsupervised learning [1], [2], [3], [4], [5], [6], [7]. In particular, the supervised learning requires manual labeling of a large amount of data, which needs to spend a lot of manpower and financial resources. Therefore, much work has proposed methods based on weak supervision and unsupervised learning to solve the problem of manual labeling.

Over the years, there are many traditional relation extraction methods [1], [8], [9]. Subsequently, some problems have also appeared, such as manually labeling a large amount of data, manually constructing Part-of-Speech Tagging or Dependency Analysis, and low accuracy. Fortunately, with the rapid development of deep learning technologies, researchers widely use deep learning methods to solve the relation extraction problems. Various methods by using Convolutional Neural Networks (CNN) models or Recurrent Neural Network (RNN) models have been proposed successively (see [10], [11], [12], [13], [14] in detail). Especially, the accuracy of relation extraction after the pre-trained language model BERT proposed by Devlin et al. [15] in 2018 has been further improved. Since the pre-trained language models can get more lexical,

[1] Corresponding Author Fu Zhang is with the School of Computer Science and Engineering, Northeastern University, Shenyang, China, PhD, Associate Professor, email: zhangfu@cse.neu.edu.cn

syntactic and semantic features, the relation extraction methods based on pre-trained language models have higher accuracy than the relation extraction methods based on CNN or RNN models. In particular, the mainstream pre-trained models based on Transformer (including BERT, GPT-2, Transformer-XL, XLNet, ROBERTa, ALBERT, etc.) almost occupy the relatively top positions in various of NLP tasks [16], [17], [18].

There are some surveys of relation extraction methods [19], [20], [21], [22], [23], [24]. Although up to now a huge number of relation extraction methods based on pre-trained language models (e.g., BERT [15]) have been proposed in the literature, to the best of our knowledge, detailed and in-depth reviews on these studies are scarce. In this paper, we provide a full up-to-date overview of the current state of the art in the existing relation extraction methods based on pre-trained language models. Regarding all the existing approaches, we make more detailed and in-depth comparisons and discussions, and we classify the existing approaches into supervised learning and distant supervision. In addition, the directions for future research and some comparisons and analyses are discussed in our whole survey. The survey may help readers understand and catch some key techniques about the issue, and identify some future research directions.

2. Development and Classification of Relation Extraction

In 1998, the last Message Understanding Conference (MUC) funded by the Defense Advanced Research Project Agency (DARPA) introduced the entity relation extraction task for the first time. Template Relation in MUC is the earliest description of entity relations [25], [26].

In 1999, the National Institute of Standards and Technology (NIST) organized an Automatic Content Extraction (ACE) evaluation, and one of the important evaluation tasks was entity relation extraction. The ACE entity relation corpus specifies 7 categories of entities, including people, organizations, facilities, premises, geopolitical entities, vehicles, and weapons, each of which is divided into multiple subcategories [23]. Since 2009, ACE has been included in Text Analysis Conference (TAC) and has become a major component of Knowledge Base Population (KBP) [27].

The entity relation extraction involved in MUC and ACE evaluation meetings is limited to a few types of entity relations between named entities (including person names, place names, organizational names, etc.), such as employment relations, geographical relations, person-society organizational relations, etc. SemEval (Semantic Evaluation) is another important evaluation conference in the field of information extraction after MUC and ACE [23]. This conference attracted a large number of institutions and research institutions to participate in the evaluation. SemEval-2007 evaluation task 4 [28] defines the entity relation between 7 common nouns or noun phrases, but the English corpus it provides is small. Subsequently, SemEval-2010 evaluation task 8 [29] enriched and perfected it, and expanded the entity relation types to 9 types, namely: Cause-Effect, Instrument-Agency, Product-Producer, Content-Container, Entity-Origin, Entity-Destination, Component-Whole, Member-Collection, Message-Topic. Considering the order of the entity pairs in the sentence instance, the "Other" class is introduced to describe instances that do not belong to the aforementioned relation types, and a total of 19 entity relations are generated [23].

Relation extraction methods are mainly divided into traditional methods based on machine learning and deep learning methods based on neural networks. The traditional methods based on machine learning are mainly divided into rule matching, supervised learning, weak supervised learning and unsupervised learning. Deep learning methods based on neural networks are mainly divided into supervised learning and distant supervision. Among them, supervised learning mainly adopts pipeline method and joint learning method. Figure 1 describes the classifications and methods of relation extraction.

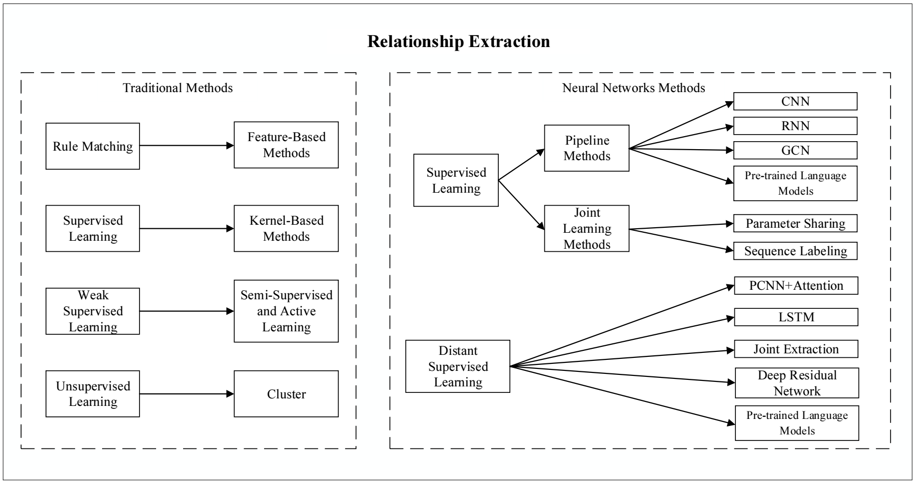

Figure 1. The classifications and methods of relation extraction.

3. Relation Extraction Datasets and Evaluation Criteria

In recent years, the standard datasets used for the evaluation of relation extraction are mainly (also see [24]): SemEval-2010 Task 8 [29], ACE2004 [30], ACE2005 [31], TACRED [32], CoNLL04 [33], FewRel [34], NYT [35], ADE Corpus [36], WebNLG [37], SciERC [38], and DocRED [39]. The relation extraction in the supervised learning mainly uses MUC [26], ACE2004 [30], ACE2005 [31], SemEval-2010 Task 8 [29], TACRED [32], and FewRel [34]. The relation extraction in the distant supervision mainly uses NYT [35] and DocRED [39].

In the supervised methods setting, relation extraction is expressed as a classification task and hence, metrics like Precision, Recall and F-Measure are used for performance evaluation [24]. In detail, Precision and Recall are evaluated based on *TP* (True Positive), *TN* (True Negative), *FP* (False Positive) and *FN* (False Negative). Suppose that *TP* is the number of correctly predicted relation instances, *TP* + *FP* is the total number of predicted relation instances, and *TP* + *FN* is the total number of relation instances, then these metrics are defined as follows:

$$\text{Precision } P = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall } R = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F} - \text{Measure } F1 = \frac{2\,P\,R}{P + R} \tag{3}$$

4. Relation Extraction by Supervised Learning based on Pre-trained Language Models

Pre-trained language models such as BERT in supervised learning relation extraction methods have deeper depths and can capture longer distance information than previous CNN-based or RNN-based supervised learning relation extraction methods. Researchers have further improved the accuracy of relation extraction by changing BERT's coding method, input-output structure, combining other models, and using knowledge base or knowledge base information. Table 1 summarizes and compares several supervised learning relation extraction methods based on BERT or pre-trained language models.

Table 1. Comparison of different supervised learning relation extraction models.

Model	*Technology*	*Use External Knowledge Base*	*Dataset*	*F1 Score*
R-BERT [40]	BERT	No	SemEval-2010 Task 8	89.25
BERT$_{EM}$ + MTB [41]	Matching the Blanks + BERT	Using Matching the Blanks	SemEval-2010 Task 8	89.5
			TACRED	71.5
			KBP37	69.3
EPGNN [42]	BERT+GCN	No	SemEval-2010 Task 8	90.2
			ACE2005	77.1
Know-Bert-W+W [43]	BERT+KAR	Wikipedia and WordNet	TACRED	71.5
			SemEval-2010 Task 8	89.1
Entity-Aware BERT [44]	BERT + Entity-Aware + Self-Attention	No	SemEval-2018 Task 7	83.9
			SemEval-2010 Task 8	89.0
TRE [45]	Pre-trained Transformer + Byte Pair Encoding	No	TACRED	67.4
			SemEval-2010 Task 8	87.1
ERNIE [46]	BERT+KG Embedding + TransE [47]	Knowledge Graph	TACRED	67.97
			FewRel	88.32
BERT-LSTM-base [48]	BERT + BiLSTM + WordPiece tokenizer [49]	No	TACRED	67.8
SpERT [50]	BERT + Span Classification+ Span filtering	No	ADE	78.84
		No	CoNLL04	71.47
		SciBERT [51] as a sentence encoder	SciERC	50.84
DYGIE++ [52]	BERT + Span Enumeration + Span Graph Propagation	No	ACE2005	63.4
			SciERC	48.4
			WLPC [53]	65.9

Wu et al. [40] proposed an R-BERT model for relation extraction tasks in 2019. In order to strengthen the BERT model to obtain the relation information between the entity e_1 and the entity e_2, they add the head entities and tail entities with symbols "$" and "#", respectively. There is a sentence s and two entities e_1 and e_2, and they suppose H is BERT's output and the final hidden state. After BERT, they get the vectors H_i to H_j and H_k to H_m, which are the final hidden vectors for the two entities e_1 and e_2. In order to obtain the vector representations for the entity e_1 and the entity e_2, they use the average operation. They add an activation operation and a fully connected layer to H_i to H_j and H_k to H_m, and get the output H_1' and H_2' for e_1 and e_2. They concatenate H_0' (an activation operation and a fully connected layer to the first token '[CLS]'), H_1', H_2' and then add a fully connected layer and a Softmax layer.

Further, the model is evaluated based on the performance metrics $F1$. In the following we show the calculation process of $F1$ score through examples (e.g., the relation Component-Whole(e_2, e_1)) from the training file in the SemEval-2010 Task 8 dataset. In detail, given a snapshot of few rows of the dataset:

sentence$_1$: The system as described above has its greatest application in an arrayed <e_1>configuration</e_1> of antenna <e_2>elements</e_2>.
sentence$_2$: The girl showed a photo of apple <e_1>tree</e_1> <e_2>blossom</e_2> on a fruit tree in the Central Valley.
sentence$_3$: The <e_1>provinces</e_1> are divided into <e_2>counties</e_2> (Shahrestan), and subdivided into districts (Bakhsh) and sub-districts (Dehestan).

If the model predicts the relation in the first sentence *sentence*$_1$ is Component-Whole(e_2, e_1), the second sentence *sentence*$_2$ is Other and the third sentence *sentence*$_3$ is None (i.e., the model doesn't predict the relation in the third sentence *sentence*$_3$). In this case, we know the predicted relation in the first sentence *sentence*$_1$ is right, the second and third sentences are wrong. As we know, *TP* is a number, which means the relation in the sentence is Component-Whole(e_2, e_1) and the predicted relation in the sentence is also Component-Whole(e_2, e_1); *FP* is a number, which means the relation in the sentence isn't Component-Whole(e_2, e_1) but the predicted relation in the sentence is Component-Whole(e_2, e_1); *FN* is a number, which means the relation in the sentence is Component-Whole(e_2, e_1) but the predicted relation in the sentence isn't Component-Whole(e_2, e_1). Then we can get *TP* is 1 easily. Because all the relations in three sentences are Component-Whole(e_2, e_1), we can get *FP* is 0. Because the predicted relations in the sentence *sentence*$_2$ and *sentence*$_3$ are wrong, we can get *FN* is 2. Eventually, we can calculate *Precision* = 1/1 = 100%, *Recall* = 1/3 = 33.3% and *F*1 score = 50%. Finally, the R-BERT model [40] achieved *F*1 score at 89.25 on SemEval-2010 Task 8. Figure 2 shows the architecture of the R-BERT model.

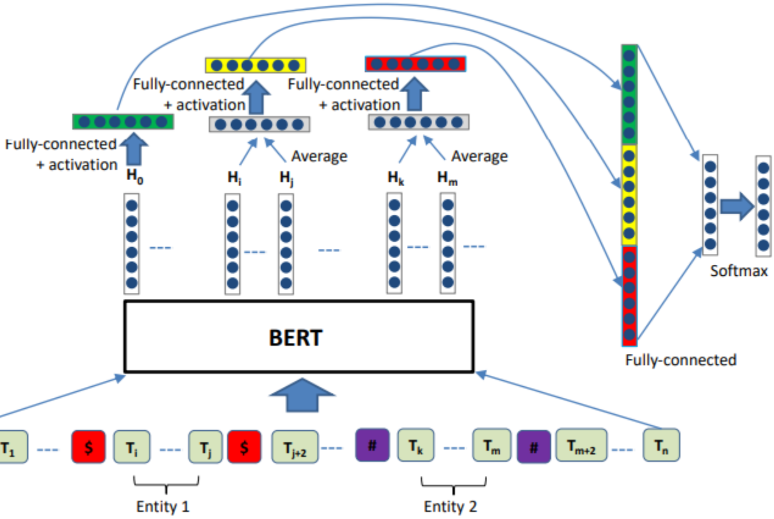

Figure 2. The architecture of the R-BERT model.

Soares et al. [41] proposed a BERT$_{EM}$ + MTB model for relation extraction tasks in 2019. They first proposed the task of MTB (Matching the Blanks), assuming that if two sentences contain the same entity pair, their relation representation should be as similar as possible, otherwise the similarity should be as low as possible. That is, *r* and *r'* representing similar relations, and their inner products $f_\theta(r) \cdot f_\theta(r')$ should be high. If input two sentences into the model and get the relation representation, according to the above assumption, the model only needs the information of the entities in the sentence (comparing whether the entity pairs are the same) to minimize the error. Therefore, they replace the entities in the sentence with the special mark "[BLANK]" according to a certain probability ($\alpha = 0.7$), so that the $\mathrm{BERT_{EM}}$ + MTB model models the context information in the sentence except the entity. The loss of the pre-trained model is the loss in the Bert Masked Language Model and the loss of the similarity of the relation. The Matching the Blanks pre-trained dataset was built from Wikipedia. The model is initialized with the parameters of Bert-Large, pre-trained on Matching the Blanks, and fine-tuned on the specific relation extraction task. Matching the Blanks task can complete the relation extraction task without a training set, and it performs well on small datasets.

After presenting the MTB, the author discussed different input and output methods of BERT. The input method is how to specify the position of two entities in the input. There are three types, which are STANDARD: standard input, without specifying the position of the entity; POSITIONAL EMB: position embedding, set the segment type of $entity_1$ and $entity_2$ tokens to 1 and 2; ENTITY MARKER: Entity mark, use the special mark entity position on both sides of $entity_1$ and $entity_2$. The output method is how to get the relational representation from the output of the last layer of BERT. There are three types, which are [CLS]: [CLS] token is used as the relational representation; MENTION POOL: use max pooling on the token representations of the two entities, and then concatenate to obtain the relation representation; ENTITY START: use a special token at the start position of two entities, spliced together as a relation representation. The paper tested the performance of different structures on SemEval-2010 Task8, KBP37, TACRED, and FewRel. They found that the ENEITY MARKER input method and ENTITY START output method performed best on all datasets. They achieved F1 score at 89.5 on SemEval-2010 Task 8 and F1 score at 71.5 on TACRED and F1 score at 69.3 on KBP37.

Zhao et al. [42] proposed an EPGNN model for relation extraction tasks in 2019. They noticed that the relation between entity pairs in a sentence can be indicated by other sentences containing the same entity pair, so the dependency relation between entity pairs in relation extraction also needs to be modeled. Therefore, they proposed the concept of entity pairs graph and used Graph Convolutional Network (GCN) to combine the semantic features and graph topological features. The BERT model was used to encode the context information.

In the model construction, they used the method of Wu et al. [40]. In order to make the BERT model capture the relation information between the two entities, the head entity and the tail entity were added with the symbols "$" and "#". Then, the sentence is encoded and input into the BERT model. The "[CLS]" at the beginning of the sentence and the hidden layer vector corresponding to the two entities are averaged and fully connected as the semantic feature of the sentence. After the average is connected, it is input to the multi-layer graph convolution network to obtain the graph topology. Finally, the semantic features of the sentences are connected to the graph topology features and input to the Softmax layer to obtain the relation extraction results. Figure 3 shows the architecture of EPGNN model. They achieved F1 score at 90.2 on SemEval-2010 Task 8 and F1 score at 77.1 on ACE2005.

Peters et al. [43] proposed the Know-Bert-W + W model for relation extraction tasks in 2019. For incorporate the structured information in the Knowledge Base into a large-scale pre-trained model, they proposed a Knowledge Attention and

Recontextualization (KAR) component. Know-Bert model integrates knowledge bases (KB) into BERT through KAR. The author proposes three KnowBert models: KnowBert-Wiki, KnowBert-WordNet, and KnowBert-W + W (including Wikipedia and WordNet). Among them, KnowBert-W + W adds the Wikipedia knowledge base to the 10 and 11 layers of BERT (Base), and WordNet to the 11 and 12 layers of BERT (Base). They achieved F1 score at 71.5 on TACRED and F1 score at 89.1 on SemEval-2010 Task 8.

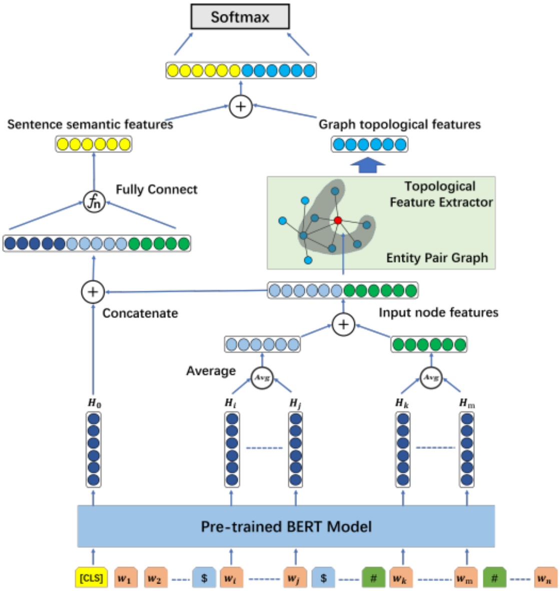

Figure 3. The architecture of EPGNN model.

Wang et al. [44] proposed the Entity-Aware BERT model for relation extraction tasks in 2019. They found that the existing Multiple Entity-Relations (MER) extraction uses a variant of Single Relation Extraction (SRE), which requires Multiple-Pass encoding when inputting the paragraph, resulting in costly calculations. It is expensive and difficult to handle long sentences. They proposed a One-Pass encoding method to solve this problem.

This method is based on BERT with a structured prediction layer to enable BERT to predict multiple relations using a single encoding, and has an entity-aware self-attention mechanism that can inject relation information with entities in hidden state. The key is to use the relative distance between words and entities to encode the location information of each entity. This information is propagated in different layers through attention computation. They achieved F1 score at 83.9 on SemEval-2018 Task 7 and F1 score at 89.0 on SemEval-2010 Task 8.

Alt et al. [45] proposed a TRE model for relation extraction tasks in 2019. Because the traditional method of relation extractions based on semantic and syntactic features requires a large amount of labeled data, which limits the generalization ability of the model, the author proposes a TRE (Transformer for Relation Extraction) model, which uses a pre-trained deep language model to capture the relation between entities. Through the unsupervised pre-training process, the model can learn implicit semantic features of sentences. They achieved F1 score at 67.4 on TACRED and F1 score at 87.1 on the SemEval-2010 Task 8.

Zhang et al. [46] proposed the ERNIE model for relation extraction tasks in 2019. They found that the previous researchers rarely thought of combining information in the Knowledge Graph (KG) when solving the relation extraction task. Therefore, they used large-scale corpora and knowledge maps to train Enhanced Language Representation with Informative Entities (ERNIE) models.

In order to solve the problem of structured knowledge coding and hybrid information fusion, they adopted two methods when constructing the ERNIE model. The first is to identify named entities in the text and match the corresponding named entities in the Knowledge Graph. They don't use the KG's graph-based information directly. They use algorithms like TransE [47] to encode the graph structure of KGs and make the informative entity embeddings as ERNIE's input. They use the MLM (Masked Language Model) and the Next Sentence Prediction as the pre-trained objectives like BERT. In addition, a method of randomly masking named entities in the input text and selecting appropriate named entities from the Knowledge Graph through ERNIE and using both context and Knowledge Graph information are also designed.

Shi et al. [48] proposed a BERT-LSTM-base model for relation extraction tasks in 2019. They found that most existing relation extraction methods rely on lexical and syntactic features, such as Part-of-Speech Tags, Syntactic Trees, Dependency Trees, and Global Decoding Constraints. Although these features improve the accuracy of relation extraction, these features are not suitable for every language, can't improve the robustness of the model, and even reduce the accuracy. Therefore, the author proposes a method that does not use external features and only uses a simple BERT model for relation extraction.

The input sentence is first converted into a "[CLS] sentence [SEP] subject [SEP] object [SEP]" form. To prevent overfitting, replace the subject and object with other symbols such as "[CLS] [S-PER] was born in [O-LOC] [SEP] Obama [SEP] Honolulu [SEP]". The input is then tokenized by the WordPiece tokenizer [49] and fed into the BERT encoder. The position vector is merged with the context vector to input a one-layer BiLSTM. The final hidden states in each direction of the BiLSTM are used for prediction with a one-hidden-layer Multilayer Perceptron (MLP). They achieved F1 score at 67.8 on the TACRED.

Eberts et al. [50] proposed a SpERT model for relation extraction tasks in 2019. They proposed a span-based joint learning model with BERT as the core, which can simultaneously extract entities and relations in sentences. Unlike traditional BIO or BILOU labels, the span-based model can identify overlapping entities. The author found that there are three aspects that help the model improve performance. The first is the negative samples in the same sentence, the second is the localized context representation, and the third is the fine-tuning of the pre-trained model. In addition, the model proposed by the author is also lightweight.

Wadden et al. [52] proposed the DYGIE ++ model for named entity recognition tasks, relation extraction tasks, event extraction tasks, and coreference resolution tasks. The model consists of 4 parts and the structure is shown in the Figure 4. The first part is Token Encoding. DYGIE++ uses BERT for token representations using a "sliding window" approach, feeding each sentence to BERT together with a size-*L* neighborhood of surrounding sentences. The second part is Span Enumeration. Spans of text are enumerated and constructed by concatenating the tokens representing their left and right endpoints, together with a learned span width embedding. The third part is Span Graph Propagation. A graph structure is generated dynamically based on the model's current best guess at the relations present among the spans in the document.

The fourth part is multi-task classification. The re-contextualized representations are input to scoring functions which make predictions for each of the end tasks. They use a two-layer feedforward neural net (FFNN) as the scoring function. Their F1 score on the relation extraction subtasks in ACE05, SciERC, and WLPC reached 63.4, 48.4, and 65.9, respectively.

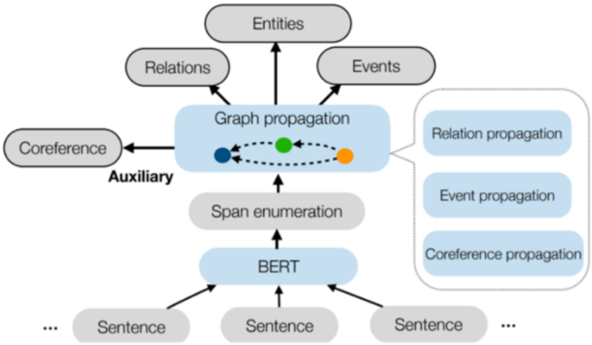

Figure 4. The architecture of DYGIE ++ model.

5. Relation extraction by distant supervision based on pre-trained language models

Distant supervision is a common method in relation extraction. This method was proposed by Mintz et al. [54] on ACL2009, and its main assumption is: if there is a relation between two entities, all sentences containing these two entities may show this relation. Riedel et al. [35] argue that this assumption is too strong, and propose at-least-once assumption: If there is a relation between two entities, at least one of all sentences containing the two entities expresses the relation. Based on this, Riedel et al. [35] modeled distant supervision as a multi-instance learning problem, and aggregated all sentences containing the same entity pair into a bag, and classified these bags. Hoffmann et al. [55] and Surdeanu et al. [56] observed that there may be more than one relation on an entity pair, so a multi-label learning method was added based on multi-instance learning. The method of distant supervision relation extraction requires aligning an unlabeled corpus to a known knowledge base. The most commonly used dataset is the NYT formed by Riedel et al. [57] aligning the New York Times with Freebase. Table 2 summarizes and compares several distant supervision relation extraction methods based on BERT or other pre-trained language models.

Table 2. Comparing different distant supervision relation extraction models.

Model	*Technology*	*Dataset*	*F1 Score*
HBT [58]	BERT + HBT	NYT	87.5
		WebNLG	88.8
BERT-Two-Step [59]	BERT +BiLinear	DocRED	53.92
REDN [60]	BERT + relation computing layer + Sigmoid classifier	SemEval-2010 Task 8	91.0
		NYT	89.8
		WebNLG	96.3

Wei et al. [58] proposed an HBT model for relation extraction tasks in 2019. They found that most of the previous work did not solve the overlapping triple problem. This problem brings challenges to traditional sequence labeling methods, and it also brings difficulties to relation extraction methods, because previous works considered that an entity has at most one relation to each other. Zeng et al. [61] and Fu et al. [62] both noticed this problem and proposed some improvements. However, there are still shortcomings. They all treat all relations as independent labels and assign them to entity pairs, which makes the relation extraction task complicated. The author proposes a brand-new method for this problem, that is, to find the most likely object and relation based on the identified subject. They implemented this method in an End-to-End Hierarchical Binary Tagging (HBT) framework. The architecture of the model is shown in Figure 5.

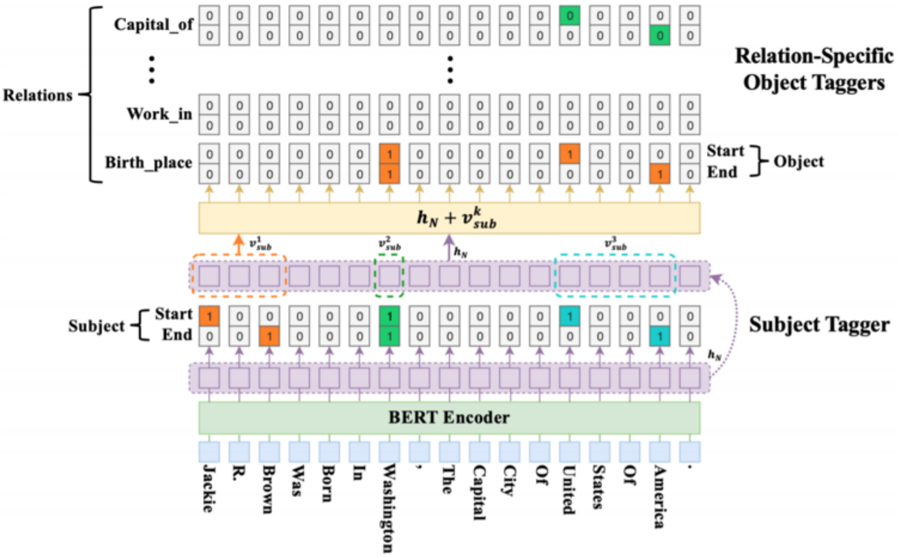

Figure 5. The architecture of HBT model.

Wang et al. [59] proposed a BERT-Two-Step model in 2019, and made progress in document-level relation extraction. They found that most of previous work focused on sentence-level data and did not focus on document-level data. A new document-level dataset DocRED based on distant supervision was released in 2019 [63], filling this gap. They found that using two steps can improve the accuracy of the model. The first step is to predict whether there is a relation between a pair of entities, and the second step is to predict the relation based on a given entity pair.

Li et al. [60] pointed out that the original standard tasks of pre-trained language model do not include the relation extraction and thus proposed a new downstream architecture of PLMs (Pre-trained Language Models) to deal with supervised relation extraction problems. Their works attempt to leverage the power of the pre-trained language models and promote the accuracy of relation extraction tasks. To achieve this goal, they implemented three improvements. First, they use BERT as encoder to extract head and tail entities embeddings from two layers. Second, in a sequence's each token, they calculate a parameterized asymmetric kernel inner-product matrix via all the head and tail embeddings. Third, they give up using Softmax classifier and choose Sigmoid classifier. Besides, for each entity pair they use the average probability of token pairs. It is regarded as the final probability for the entity pair which has a certain relation. Their method reached F1 score at 91.0 on SemEval-2010 Task 8 and F1 score at 89.8 on NYT and F1 score at 96.3 on WebNLG.

6. Results and Discussions

This section makes some results and discussions about the different models as mentioned in the previous Sections 4 and 5. Through comparison as shown in Table 3, it is concluded that each model uses different bright-spot techniques to obtain performance improvement, so that the reader can compare the advantages of each model.

Table 3. Comparisons of different relation extraction models.

Model	*The main bright spots*
R-BERT [40]	The model marks entities with $ and # to enhance BERT's ability to capture entities.
BERT_{EM}+MTB [41]	Matching the Blanks improves the accuracy of relation extraction. Using ENEITY MARKER input and ENTITY START output further improves the performance of BERT.
EPGNN [42]	The model marks entities with $ and # to enhance BERT's ability to capture entities. The model uses GCN to extract the topological structure of the entity pair graph, which is combined with the semantic information of the sentence to further improve the accuracy of relation extraction.
Know-Bert-W+W [43]	The Know-Bert model integrates knowledge bases (KB) into BERT through KAR, so that the model can combine information from external knowledge bases, thereby improving the accuracy of relation extraction.
Entity-Aware BERT [44]	The model uses One-Pass encoding to solve the problem of Multiple Entity-Relations (MER) extraction. Entity-Aware and Self-Attention mechanisms can inject relation information for multiple entities in each hidden state, thereby improving the accuracy of MER extraction.
TRE [45]	The model uses multiple layers of Transformer to encode the input sequence. In terms of input expression, the input sentence is encoded using Byte Pair Encoding (BPE), which improves the model's ability to learn hidden semantic features.
ERNIE [46]	The model incorporates Knowledge Graph (KG) information during the training process of the multi-layer Transformer model, so that it can learn structured knowledge from KG, thereby improving the accuracy of relation extraction.
BERT-LSTM-base [48]	BERT combined with BiLSTM improves the robustness of the model.
SpERT [50]	Based on the Span-Based joint learning model with BERT as the core, overlapping entities can be extracted. After the sentences are encoded by the BERT layer, the overlapping entities are identified by the Span Classification layer and the Span Filtering layer, and the relation between the entities is obtained by the Relation Classification layer.
DYGIE ++ [52]	The model solves a variety of tasks by encoding sentences with BERT, spanning enumeration and span graph propagation.
HBT [58]	In order to solve the overlapping triple problem, the model uses a Hierarchical Binary Tagging (HBT) framework to extract the subjects and objects in the overlapping triples and the relation between them.
BERT-Two-Step [59]	To solve the problem of document-level data extraction, the model uses two steps to improve the accuracy of the model. The first step is to predict whether there is a relation between a pair of entities, and the second step is to predict relation based on a given entity pair.
REDN [60]	The method implemented three improvements to establish an effective downstream model that is competent for the relation extraction tasks. The first is using BERT as encoder. The second is using a parameterized asymmetric kernel inner product matrix. The third is using Sigmoid classifier instead of the Softmax classifier.

7. Conclusions and Future Research Directions

In this paper we provided a review of relation extraction based on pre-trained language models. The survey in this paper may help readers know and catch some key techniques about this issue and may identify some future research directions. After the release of pre-trained language models such as BERT, they have shown great potential in relation extraction, which has greatly improved the accuracy in a short time. However, the researches on relation extraction are still in a developing stage. The following issues may be important in order for relation extraction technologies to be more widely adoptable in many application domains:

- Document-level relation extraction: Most of the current relation extraction tasks focus on extracting relations between entity pairs in a sentence, but rarely focus on document-level data. The combination of coreference resolution and relation extraction is a solution to this problem. Peng et al. [64] proposed a general relation extraction framework based on Graph LSTM in 2017, which can be easily extended to cross-sentence N-gram relation extraction. Hence the relation extraction based on graph structure is the solution to this problem.

- Mislabeling of distant supervision: Mislabeling of distant supervision is the most important factor affecting the accuracy of relation extraction [24]. At present, this problem is mainly solved by three methods: The first is an instance selection method. By using the Attention mechanism to assign different weights to sentences of different confidence levels or using multi-instance learning to label test bags, the weight of mislabeled instances noise is reduced. The second method used by Fan et al. [65] and Luo et al. [66] restore the true labels by modeling the process of noise generation. The third is that Qin et al. [67] introduced deep reinforcement learning into distant supervision, and put example sentences with no target relation into the negative set. The mislabeling of distant supervision is an important problem.

- Relation extraction uses knowledge base and Knowledge Graph information: It is one of the difficult points to effectively use the knowledge base and Knowledge Graph information to make a large amount of structured knowledge to help the relation extraction. Peters et al. [43] used the KAR layer to incorporate the knowledge base information, and Zhang et al. [46] used the TransE [47] algorithm to match the entities in the Knowledge Graph with the entities in the training corpus when the pre-trained model was trained. And their model obtained structured information by using the method above.

- Interpretability of BERT model: Jawahar et al. [68] and Bouaoui et al. [69] respectively explored BERT's deep-level representation learning and induction of relation knowledge from BERT in 2019. This is a very necessary thing. First, it can help us understand the limitations of BERT more clearly, so as to improve BERT or figure out its application scope. Second, it helps to explore the interpretability of BERT. Future research can start from this aspect and further explore the interpretability of deep neural networks.

- Identify the relation of overlapping entities: The relation of overlapping entities recognition has always been an obstacle to further improving the accuracy of relation extraction. Zheng et al. [70], Katiyar et al. [71], Zeng et al. [61], and Fu et al. [62] have all done some research on this problem, but none have achieved particularly good results. The method in [58] may well solve this problem, but there is still room for further research on cross-sentence and document-level relation extraction.

Acknowledgments

The authors thank the anonymous referees for their valuable comments and suggestions. The work is supported by the National Natural Science Foundation of China (61672139).

References

[1] Bach N, Badaskar S. A review of relation extraction. Literature review for Language and Statistics II. 2007; 2: 1-15.
[2] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. ACL on Interactive poster and demonstration sessions, 2004. p. 22-25.
[3] Zhao S, Grishman R. Extracting relations with integrated information using kernel methods. ACL, 2005. p. 419-426.
[4] Jiang J, Zhai C X. A systematic exploration of the feature space for relation extraction. NAACL-HLT, 2007. p. 113-120.
[5] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction. Journal of machine learning research. 2003; 3(2): 1083-1106.
[6] Culotta A, Sorensen J. Dependency tree kernels for relation extraction. ACL, 2004. p. 423-429.
[7] Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction. HLT/EMNLP, 2005. p. 724-731.
[8] Alisa Smirnova and Philippe Cudré-Mauroux. Relation extraction using distant supervision: a survey. ACM Computing Surveys. 2019; 51(5): 1-35.
[9] Konstantinova N. Review of relation extraction methods: what is new out there?. International Conference on Analysis of Images, Social Networks and Texts, 2014. p. 15-28.
[10] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces. EMNLP-CoNLL, 2012. p. 1201-1211.
[11] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network. COLING, 2014. p. 2335-2344.
[12] Xu K, Feng Y, Huang S, et al. Semantic relation classification via convolutional neural networks with simple negative sampling. EMNLP, 2015. p. 536-540.
[13] Vu N T, Adel H, Gupta P, et al. Combining recurrent and convolutional neural networks for relation classification. NAACL-HLT, 2016. p. 534-539.
[14] Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths. EMNLP, 2015. p. 1785-1794.
[15] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT, 2019. p. 4171-4186.
[16] SuperGLUE. https://super.gluebenchmark.com.
[17] NLP-progress. https://nlpprogress.com.
[18] AI2. https://leaderboard.allenai.org.
[19] Kumar S. A survey of deep learning methods for relation extraction. arXiv preprint arXiv:1705.03645, 2017.
[20] Li A, Wang X, Wang W, et al. A survey of relation extraction of knowledge graphs. Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, 2019. p. 52-66.
[21] Goyal K, Bhattacharyya P. Literature survey on relation extraction and relational learning. Indian Institute of Technology. 2017; 1-19.
[22] Zhang Q, Chen M, Liu L. A review on entity relation extraction. Second International Conference on Mechanical, Control and Computer Engineering, 2017. p. 178-183.
[23] Liu S Y, Li Y C, Guo Z G, Wang B, Chen G. Review of entity relation extraction. Journal of Information Engineering University. 2016; 17(05): 541-547.
[24] E HH, Zhang WJ, et al. Survey of entity relationship extraction based on deep learning. Journal of Software. 2019; 30(6): 1793-1818.
[25] Chinchor N, Marsh E. Muc-7 information extraction task definition. Proceeding of the seventh message understanding conference (MUC-7), 1998. p. 359-367.
[26] Chinchor N A. Overview of muc-7/met-2. Science Applications International Corp San Diego CA, 1998.

[27] McNamee P, Dang H T, Simpson H, et al. An evaluation of technologies for knowledge base population. LREC, 2010. p. 369-372.

[28] Girju R, Nakov P, Nastase V, et al. Semeval-2007 task 04: Classification of semantic relations between nominals. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval), 2007. p. 13-18.

[29] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way Classification of Semantic Relations between Pairs of Nominals. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions, 2009. p. 94-99.

[30] ACE 2004. [EB/OL]. https://catalog.ldc.upenn.edu/LDC2005T09.

[31] ACE 2005. https://catalog.ldc.upenn.edu/LDC2006T06.

[32] TACRED. [EB/OL]. https://nlp.stanford.edu/projects/tacred/.

[33] Roth D, Yih W. A linear programming formulation for global inference in natural language tasks. Illinois univ at urbana-champaign dept of computer science, 2004.

[34] Han X, Zhu H, Yu P, et al. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. EMNLP, 2018. p. 4803-4809.

[35] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2010. p. 148-163.

[36] Gurulingappa H, Rajput A M, Roberts A, et al. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of biomedical informatics. 2012; 45(5): 885-892.

[37] Gardent C, Shimorina A, Narayan S, et al. Creating training corpora for nlg micro-planning. ACL, 2017. p. 179-188.

[38] Luan Y, He L, Ostendorf M, et al. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. EMNLP, 2018. p. 3219-3232.

[39] Yao Y, Ye D, Li P, et al. Docred: A large-scale document-level relation extraction dataset. ACL, 2019. p. 764-777.

[40] Wu S, He Y. Enriching pre-trained language model with entity information for relation classification. CIKM, 2019. p. 2361-2364.

[41] Soares L B, FitzGerald N, Ling J, et al. Matching the blanks: distributional similarity for relation learning. ACL, 2019. p. 2895-2905.

[42] Zhao Y, Wan H, Gao J, et al. Improving relation classification by entity pair graph. Asian Conference on Machine Learning, 2019. p. 1156-1171.

[43] Peters M E, Neumann M, Logan I V, et al. Knowledge enhanced contextual word representations. EMNLP-IJCNLP, 2019. p. 43-54.

[44] Wang H, Tan M, Yu M, et al. Extracting multiple-relations in one-pass with pre-trained transformers. ACL, 2019. p. 1371-1377.

[45] Alt C, Hübner M, Hennig L. Improving relation extraction by pre-trained language representations. arXiv preprint arXiv:1906.03088, 2019.

[46] Zhang Z, Han X, Liu Z, et al. ERNIE: enhanced language representation with informative entities. ACL, 2019. p. 1441-1451.

[47] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 2013. p. 2787-2795.

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

[49] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. ACL, 2016. p. 1715-1725.

[50] Eberts M, Ulges A. Span-based joint entity and relation extraction with transformer pre-training. arXiv preprint arXiv:1909.07755, 2019.

[51] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. EMNLP-IJCNLP, 2019. p. 3606-3611.

[52] Wadden D, Wennberg U, Luan Y, et al. Entity, relation, and event extraction with contextualized span representations. EMNLP-IJCNLP, 2019. p. 5788-5793.

[53] Kulkarni C, Xu W, et al. An annotated corpus for machine reading of instructions in wet lab protocols. NAACL, 2018. p. 97-106.

[54] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data. ACL and AFNLP, 2009. p. 1003-1011.

[55] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations. ACL, 2011. p. 541-550.

[56] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction. EMNLP-CoNLL, 2012. p. 455-465.

[57] Sandhaus E. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, 2008.

[58] Wei, Zhepei, et al. A novel hierarchical binary tagging framework for joint extraction of entities and relations. arXiv preprint arXiv:1909.03227, 2019.
[59] Wang H, Focke C, Sylvester R, et al. Fine-tune Bert for DocRED with two-step process. arXiv preprint arXiv:1909.11898, 2019.
[60] Li C, Tian Y. Downstream model design of pre-trained language model for relation extraction task. arXiv preprint arXiv:2004.03786, 2020.
[61] Zeng X, Zeng D, et al. Extracting relational facts by an end-to-end neural model with copy mechanism. ACL, 2018. p. 506-514.
[62] Fu T J, Li P H, Ma W Y. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. ACL, 2019. p. 1409-1418.
[63] Yuan Yao, Deming Ye, Peng Li, et al. DocRED: A large-scale document-level relation extraction dataset. ACL, 2019. p. 764-777.
[64] Peng N, Poon H, Quirk C, Toutanova K, Yih WT. Cross-sentence N-ary relation extraction with graph LSTMs. Transactions of the Association for Computational Linguistics. 2017; 5: 101-115.
[65] Fan M, Zhao D, Zhou Q, et al. Distant supervision for relation extraction with matrix completion. ACL, 2014. p. 839-849.
[66] Luo B, et al. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. ACL, 2017. p. 430-439.
[67] Qin P, Xu W, Wang WY. Robust distant supervision relation extraction via deep reinforcement learning. ACL, 2018. p. 2137-2147.
[68] Jawahar G, Sagot B, Seddah D, et al. What does BERT learn about the structure of language?. ACL, 2019. p. 3651-3657.
[69] Bouraoui Z, Camacho-Collados J, Schockaert S. Inducing Relational Knowledge from BERT. AAAI, 2019. p. 1-8.
[70] Zheng S, Wang F, et al. Joint extraction of entities and relations based on a novel tagging scheme. ACL, 2017. p. 1227-1236.
[71] Katiyar A, Cardie C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. ACL, 2017. p. 917-928.