Self-Knowledge Enhanced Academic Full-Text Relation Extraction (Postprint)

Authors: Zhuo Keqiu, Shen Si, Wang Dongbo

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] Relation extraction from academic full text constitutes a key technology for constructing academic full-text knowledge graphs. Such knowledge graphs enable the structuring and intellectualization of literature, enhance researchers' efficiency in literature retrieval, analysis, and tracking of research developments, and facilitate implicit knowledge discovery through cognitive reasoning on the graph. [Method/Process] Although enhancing relation extraction with external knowledge has yielded promising results in numerous studies, domain-specific relation extraction often lacks available external knowledge. Our research demonstrates that high-confidence knowledge inherent in full text itself can be leveraged to assist full-text relation extraction. Inspired by the dual-system theory of cognitive processes (System 1 as intuitive cognition, System 2 as inferential cognition), we design a sentence-level model to acquire knowledge and obtain high-confidence knowledge via distant supervision, which is then integrated into the final classification layer of the full-text-level deep learning model. [Results/Conclusion] On the biomedical academic full-text dataset (CDR-revised), our approach achieves an 11.13% improvement in F1 score over current state-of-the-art models.

Full Text

Research on Relation Extraction from Academic Full-Text Based on Self-Owned Knowledge Enhancement

Zhuo Keqiu[1], Shen Si[2], Wang Dongbo[1]

[1]School of Information Management, Nanjing Agricultural University, Nanjing 210095
[2]School of Economics and Management, Nanjing University of Technology, Nanjing 210094

Abstract:
[**Purpose/Significance**] Relation extraction from academic full-text is a key technology for constructing academic full-text knowledge graphs. The constructed academic knowledge graph can realize the structuring and intellectualization of documents, improve the efficiency of researchers in retrieving and analyzing literature and grasping research trends, and facilitate implicit knowledge discovery through cognitive reasoning on the graph. [**Method/Process**] While enhancing relation extraction through external knowledge has achieved results in many studies, relation extraction for specific domains often lacks available external knowledge. This study found that high-confidence knowledge inherent in full-text can also assist full-text relation extraction. Inspired by the dual-system theory of cognitive processes (System 1 as intuitive cognition and System 2 as reasoning cognition), we designed a sentence-level model to acquire knowledge, obtained high-confidence knowledge through distant supervision, and integrated this high-confidence knowledge into the final classification layer of the full-text-level deep learning model. [**Result/Conclusion**] On the biomedical academic full-text dataset (CDR-revised), the F1 score improved by 11.13% compared to the current state-of-the-art model.

Keywords: academic full-text; relation extraction; self-owned knowledge enhancement; knowledge graph
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2022.07.012

1 Introduction

Relation extraction aims to assign semantic relations to entities in unstructured text. According to the scope of text processing, relation extraction can be divided into sentence-level relation extraction and full-text-level relation extraction. Sentence-level relation extraction aims to identify the relationship between two known entities within a sentence, whereas full-text-level relation extraction targets extracting relationships among multiple entities across long texts containing multiple sentences. An example of full-text-level relation extraction is shown in [Figure 1: see original paper]. Relation extraction is a key technology for knowledge graph construction. The construction of academic full-text knowledge graphs naturally depends on research into relation extraction from academic full-text. Since Google proposed knowledge graphs for search engine projects in 2012, knowledge graphs have gradually replaced the semantic web as a major research hotspot in artificial intelligence. Currently, knowledge graphs have wide applications in semantic search, intelligent question answering, knowledge engineering, data mining, and digital libraries. Combining knowledge graphs with academic full-text presents many valuable research questions.

The construction of academic full-text knowledge graphs can be roughly divided into two aspects from the perspective of whether they focus on macro-level

or micro-level academic research knowledge: macro-level knowledge graph construction of academic research achievements, and micro-level knowledge graph construction of discipline knowledge contained within achievements. Macro-level academic research achievements mainly include researcher information, publication venue information, research methods, research problems, research results, research prospects, cited research problems, cited research results, cited research methods, etc. Discipline knowledge within academic achievements mainly refers to domain knowledge. For example, in the biomedical field, a research paper [1] states that the chemical tacrolimus can induce scleroderma renal crisis, from which we can derive that the chemical tacrolimus has a certain relationship with the disease scleroderma renal crisis. The relation extraction studied in this paper targets the second scenario—extracting discipline knowledge relationships from academic research achievements.

In recent years, many methods have been proposed for relation extraction, including traditional methods dependent on manual feature engineering [2] and neural network-based models [3-4]. Neural network-based methods achieve state-of-the-art performance through end-to-end training to extract features. These neural network-based methods utilize position features to obtain entity information, specifically by providing the relative distance between each word and the two entities as model input. Recent work has applied pre-trained models (such as BERT [5]) to relation extraction. Since full-text-level relation extraction involves multiple target entities, entity marking methods that provide information for two entities at once are no longer applicable because they cannot provide information for all entities simultaneously.

To address this problem, most full-text-level relation extraction methods are based on graph models with appropriate modifications to achieve good results [6] (see [Figure 1: see original paper]). Specifically, they use words as nodes and intra-sentential and inter-sentential dependencies (dependency structures, coreference, etc.) as edges. Graph models provide a unified method for extracting features of entity pairs. Subsequent work has extended graph model approaches by improving neural network structures [7-8] or adding more types of edges [9-10]. However, entity relation extraction in full-text is extremely complex, and some entity relationships are difficult to achieve ideal results with only global-local mixed modeling. If additional knowledge can assist relation judgment, better results can be obtained.

Therefore, we propose a novel self-owned knowledge acquisition model to enhance relation extraction from academic full-text. This is the first work to enhance relation extraction from the perspective of self-owned knowledge rather than external knowledge (such as knowledge graphs), and it also serves as a validation of dual-system cognitive theory. We introduce advanced technologies such as multi-view graph models, multi-path reasoning networks, and adaptive threshold selection to ensure the accuracy of academic full-text relation extraction. Compared with existing common deep learning relation extraction methods, the extraction effect is significantly improved. Experimental results on

the biomedical academic text CDR-revised and GDA datasets demonstrate the effectiveness of the proposed method, particularly outperforming recent baseline models, thereby promoting the further improvement of key technologies for relation extraction in academic full-text knowledge graph construction and accelerating the implementation of academic full-text retrieval and knowledge construction.

2 Related Work

2.1 Evolution of Deep Learning-Based Relation Extraction Technology

Through literature content analysis, we summarize the technical evolution of deep learning-based relation extraction. As shown in [Figure 2: see original paper], the evolution of relation extraction technology is mainly reflected in word features, external knowledge, deep models, training dataset scale, extraction effect, and research fields. In terms of word features, features such as words, part-of-speech, syntactic relations, and WordNet hypernyms were initially introduced, followed by entity type identifiers and word position features. The introduction of entity types can narrow the scope of relation categories, while word position features can reflect contextual semantic information between words. Regarding external knowledge, early methods commonly used distant supervision and transfer learning, while recent years have shown a preference for knowledge graph fusion research. In fact, distant supervision, transfer learning, and knowledge graph fusion can all improve relation extraction accuracy to a certain extent, but knowledge graphs can provide more effective auxiliary information, and with the gradual improvement of domain knowledge graphs, researchers prefer knowledge graph fusion as external knowledge to improve performance.

In terms of deep models, influenced by the gradual improvement of general deep learning models, relation extraction research has continuously introduced the latest and most efficient general models to enhance extraction effects. Regarding training scale, also influenced by general pre-trained language models, relation extraction research has shifted from larger training datasets to smaller datasets to adapt to the reality of high corpus annotation costs. In terms of extraction effect, after 5-6 years of development, the F1 score has improved by about 10%, with significant results. In terms of research fields, researchers initially used easily obtainable internet texts (such as Wikipedia) as research objects. Later, as the research community grew, relation extraction research gradually blossomed in various disciplinary subfields, constructing domain-specific knowledge graphs.

2.2 Relation Extraction Related Research

Since full-text-level relation extraction is much more complex than sentence-level relation extraction, initial research mainly focused on sentence-level relation extraction [13-15]. Sentence-level relation extraction aims to detect relationships between entities in sentences. Existing sentence-level relation extraction mod-

els can be divided into two categories: sequence-based and dependency-based. Sequence-based models operate only on word sequences, which can be unidirectional or bidirectional [16-17]. Sequence-based models are relatively simple to implement but are susceptible to interference from other words in the context, making it difficult to effectively and accurately capture the semantic relationship of target entity pairs. Dependency-based models incorporate dependency trees into the model, which can theoretically avoid the interference problem of sequence-based models. However, they are difficult to implement for two main reasons: (1) dependency trees mainly rely on syntactic parsing for generation, which can easily lead to error accumulation; (2) the integration technology of dependency tree models into deep learning models is not yet mature.

L. B. Soares et al. found that entity marker methods can effectively improve the accuracy of relation extraction [18]. This entity marking approach has been widely applied in subsequent research. However, in typical article writing, it is difficult to describe relationships clearly in a single sentence, especially when descriptions contain multiple entities and relationships. Full-text-level relation extraction requires the ability to extract relationships across longer contextual sentences. Recent work has begun to explore full-text-level relation extraction. Y. Yao et al. utilized DocRED, a large-scale general dataset publicly available from Wikipedia and Wikidata, making significant progress in full-text-level relation extraction [19]. Most methods for full-text-level relation extraction are based on graph neural networks to capture semantic information between sentences [7,20-22]. F. Christopoulou et al. constructed graphs containing different granularities (sentence, mention, entity) through co-occurrence and heuristic rules, modeling the graphs without external tools [23]. S. Zeng et al. built dual graphs of different granularities to capture document-aware features and interactions between entities [24]. Z. Guo et al. proposed a refinement mechanism to aggregate multi-hop information across the entire document, and their LSR model achieved good performance in full-text-level relation extraction [25]. Graph neural networks leverage their inherent advantage of expressing dependencies between nodes through graph structures, which can partially solve the problem of sequence models being disturbed by other words in the context. However, due to the limitations of actual computing hardware, graph neural networks often need to split a complete large graph into several small graphs to adapt to mini-batch end-to-end operation mechanisms, and the related implementation technology of finding subgraphs within large graphs needs further research.

W. Xu et al. designed a discriminative reasoning network that estimates the probability distribution of different reasoning paths based on the constructed graph and the context vector of each entity pair to identify relationships between entity pairs [22]. The ATLOP model is currently known to be the most effective model on the open-source full-text corpus DocRED [26]. This model proposes an adaptive threshold technique that replaces the global threshold with a learnable threshold class. This technique eliminates the need for threshold adjustment and allows the threshold to be adjusted according to different entity pairs, thereby

obtaining better results.

Using existing knowledge graphs to guide relation extraction is another development direction [27-29]. Knowledge graphs contain a large amount of entity relationship information, which can effectively compensate for insufficient data during relation extraction training. Methods for fusing knowledge graphs into relation extraction mainly include: (1) fusion from the perspective of model features, adding entity type information to the attention mechanism to enable relation extraction models to capture text semantic features more effectively; (2) fusion from the perspective of supervised training, pre-training knowledge graphs and using knowledge graph embeddings to supervise relation extraction models, effectively reducing noise signals in current relation extraction training sets; (3) fusion from the perspective of category reasoning, exploring zero-shot learning to capture hierarchical structures between relations and correlations, enabling some relations with very little training data to be fully identified.

Q. Chen et al. used basic knowledge such as synonyms, antonyms, hypernyms, and co-hyponyms to help build a soft alignment neural network model between sentence pairs [30]. However, it can only handle a fixed number of knowledge types and requires pre-assigning values to relations before training, which limits its application in practice. Z. Wang et al. proposed a knowledge graph-enhanced natural language inference (KGNLI) model [31]. KGNLI first extracts entities such as subjects, predicates, and objects from given sentence pairs, then learns knowledge relationship representations based on a knowledge graph containing these entities as nodes. Additionally, KGNLI learns semantic relationship representations between given sentences through a bidirectional LSTM network. Finally, KGNLI combines these two representations and inputs them into a multi-layer perceptron to determine relation labels. However, the key words determining the relationship between sentence pairs are difficult to find, and KGNLI is no exception. M. E. Peters et al. proposed a Knowledge Attention and Context Reconstruction (KAR) component to improve the BERT model and enhance its ability for relation extraction [32]. However, KAR only fuses word embeddings from knowledge graphs with word embeddings from sequences, which cannot fully leverage the relational knowledge of entity pairs on knowledge graphs, thereby affecting the upstream context of entity pairs to be judged.

3 Research Methods

3.1 Framework for Academic Full-Text Relation Extraction with Self-Owned Knowledge Enhancement

Compared with non-academic texts, academic texts are more rigorous in writing, with clearer arguments and evidence, and clearer logic, resulting in more complex sentences and more logical reasoning relationships in context. Therefore, we designed a self-owned knowledge-enhanced academic full-text relation extraction model to improve the accuracy of relation extraction in academic full-text. As shown in [Figure 3: see original paper], the framework is divided into two

main parts. The left part is used to obtain sentence-level self-owned knowledge, and the right part is used to obtain full-text-level entity pair relations, with the self-owned knowledge from the left part used to guide the full-text-level relation extraction on the right part.

When performing academic full-text relation extraction, we must consider not only intra-sentential semantic logic but also inter-sentential semantic logic, with greater emphasis on logical reasoning than in non-academic texts. Based on this, we start with reasoning to solve the problem of academic full-text relation extraction (corresponding to the right half of [Figure 3: see original paper]). The method is detailed below.

3.2 Reasoning-Based Academic Full-Text Relation Extraction

Academic full-text relation extraction involves multiple types of reasoning, mainly including pattern matching, logical reasoning, coreference reasoning, and commonsense reasoning [33]. Currently, most related research uses only a single graph model to obtain low-dimensional distributed representations of entity pairs through multi-hop graph convolution and then calculates the relationship between them to complete various types of reasoning. Although this approach can achieve good reasoning results to a certain extent, it overlooks the technique that different reasoning forms require different modeling strategies. Referencing the approach of W. Xu et al., we divide reasoning into intra-sentential reasoning, logical reasoning, and coreference reasoning, and establish different reasoning paths and modeling for these three types of reasoning [22].

Typically, an entity pair contains multiple entity mentions, meaning that an entity pair can have multiple relationships through the above three reasoning methods. For the logical reasoning part, since there may be multiple entity mentions mk, there will also be multiple probabilities for this part. Our strategy is to retain the relationship with the highest probability among entity mention pairs as the possible relationship for the entity pair. Only when the relationship probability exceeds a certain threshold will the relationship be output; otherwise, no relationship is output. The method for obtaining this threshold is described below.

Generally, a global probability threshold is compared with probability Pr to finally determine the category of the entity pair. The global probability threshold is usually obtained through multiple experiments calculating the F1 score on the validation set, with the threshold determined when the F1 score is maximized. This method has two drawbacks: (1) it may not be suitable for all categories—for example, category 1 may require a probability threshold of 0.5, while category 2 may require 0.4; (2) it requires multiple runs on the validation set to obtain the global probability threshold, resulting in high time complexity.

Following the approach of W. Zhou et al., we use a learnable threshold class instead of a global threshold [27]. Specifically, during model training, we in-

troduce a virtual threshold class to separate positive and negative categories. All positive category probabilities must be higher than the virtual threshold class probability, while all negative category probabilities must be lower than the virtual threshold class probability. During inference, categories with probabilities higher than the virtual threshold class are returned as predicted labels, or if no category exceeds the virtual threshold class, the "no relation" label is returned. This technique allows the threshold to be adjusted according to different entity pair relation categories, yielding better results while reducing the time complexity of obtaining thresholds through multiple validation set runs.

We found that relying solely on the above reasoning methods to identify relationships in academic texts, while addressing the need for inter-sentential and intra-sentential reasoning, is still susceptible to interference from other words, phrases, and sentences in the context due to long-range dependencies between entities. To solve this problem, we propose to further improve academic full-text relation extraction by using an independent model with strong intra-sentential reasoning capabilities. This independent model, the self-owned knowledge acquisition model (corresponding to the left half of [Figure 3: see original paper]), is described in detail below.

3.3 Self-Owned Knowledge Acquisition Model

Inspired by the dual-system theory in cognitive science, we first attempt to obtain simple and clear entity relationships from sentences and then use these relationships to assist complex reasoning in full-text relation extraction.

Relationships in sentences are usually clear, but the automatic identification effect is greatly reduced when encountering multiple entities. How to ensure the acquired relationships have high credibility is a worthwhile research question. One approach is to train the model only on sentences with few entities and predict new sentences with few entities to obtain relationships in new sentences. Generally, this method achieves high accuracy for sentence relationships but misses much relational knowledge. The second approach builds on the first by additionally attempting to discover relationships from sentences with multiple entities, but this is more challenging and requires carefully designed extraction methods. We adopt the second approach to obtain high-confidence entity pair relationships. As shown in the left part of [Figure 3: see original paper], the process is mainly divided into three modules: a BERT module, a multi-view graph module, and a distant supervision module, to obtain entity pair relationships with high accuracy as much as possible. The entity pair relationships obtained from sentences are called self-owned knowledge and can be used for subsequent full-text relation extraction. The three modules are described in detail below.

3.3.1 BERT Module We use SciBERT [34] as the encoder to extract semantic information from sentences. SciBERT is trained on academic corpora and has only 42% vocabulary overlap with the ordinary BERT model, indicating significant differences in commonly used words between academic domain

texts and general domain texts. SciBERT has been proven to outperform the ordinary BERT model on various language tasks for academic texts. Words in sentences are input as tokens into SciBERT, and after encoding, the output is $H = \{h_0, h_1, h_2, \cdots, h_1, h\}$. Typically, h_0 is used to represent the sentence and for category label judgment. Following the method of F. Xue et al., we add a graph module layer on top of the SciBERT encoder output H, and then combine the learned graph with h_0 as input to the classifier [35]. The added graph module, the multi-view graph module, is described below.

3.3.2 Multi-View Graph Module (1) Gaussian Graph and Convolution Calculation. The BERT module outputs $H = \{h_0, h_1, \cdots, h\}$, where h_0 is derived from the [CLS] token at the beginning of the sentence and does not need to be passed into the graph module. The remaining encoded representations are input into the graph module and labeled as $V^0 = \{v_1{}^0, \cdots, v^0\}$, where N Gaussian distributions $\{N_1{}^0, \cdots, N^0\}$ are generated for each v^0 ($i = 1 \cdots m$). The expectations and variances of the Gaussian distributions are obtained through trainable neural networks. The main reasons for using a multi-view approach to generate Gaussian distributions are: (1) it can capture as many meanings of tokens as possible; (2) when the prior distribution of tokens is unknown, choosing a Gaussian distribution is a relatively safe decision, as the central limit theorem shows that the sum of many independent random variables approximately follows a Gaussian distribution, and many real distributions are themselves close to Gaussian distributions.

(2) Dynamic Temporal Pooling and Classifier Module. After each convolutional layer, a dynamic temporal pooling (DTWPool [35]) is applied. For the nth view of the graph, we first calculate the attention of each node, then use the SAGPool [36] method to filter nodes, and the remaining node set is a subset of the original node set. After L layers of pooling, we obtain L graphs $\{G_1, G_2, \cdots, G\}$, where the nodes in each graph are the union of N views. Since the length of each sentence is inconsistent, the number of nodes containing effective information in the graphs is also inconsistent, requiring a pooling mechanism to retain important node information. The solution is to introduce a loss function that supports inconsistent node numbers, minimizing the difference between G_1 and G . This approach can capture more local information to the maximum extent. The final result graph is merged from graphs at various levels. Since the number of nodes in graphs at each level differs, we only select graphs that contain the same nodes as G or are subsets of G .

For the graph output by dynamic temporal pooling, after one more max pooling, we obtain the graph vector representation. This graph vector representation can assist the [CLS] representation h_0 obtained by the BERT module encoding, making relation classification more accurate. By concatenating h_0 and the graph vector and passing them through a softmax layer, we can obtain the relation category label for an entity pair in a sentence. When a sentence contains multiple entity pairs, the above calculation is repeated, judging only one entity pair

at a time.

3.3.3 Distant Supervision Module Although the BERT model can learn basic language knowledge and improve extraction results, it still relies on labeled data. To address this issue, a method called distant supervision has been applied [37-40]. Distant supervision is also known as weak supervision. In relation extraction, distant supervision mainly uses entity pairs and relations from knowledge graphs and compares them with various available texts. When entity pairs appear simultaneously in a text, the text is considered to contain that relation. We also borrow this idea, placing the entity pairs obtained above into a search engine. When the top n search results contain entries that simultaneously include the entity pair, the relation corresponding to the entity pair is considered correct; otherwise, the entity pair relation is discarded. In experiments, top n is set to 10 and is set to 3. The reason top n is set to 10 is that the first page of search engine results usually contains 10 entries, and the information on the first page is sufficient without needing to use content from other pages. The reason is set to 3 is explained in Section 4.6 below. As shown in [Figure 4: see original paper], the entity pair "motor disorder" and "levodopa" appear simultaneously in Baidu' s search results and meet the screening requirements, so the entity pair is retained and their relation is used as knowledge to enhance subsequent full-text relation extraction.

3.3.4 Self-Owned Knowledge Enhancement When an entity pair has no relation in the full-text relation extraction result, we check the relation determined in the self-owned knowledge. If a relation exists in the self-owned knowledge, we change the full-text relation extraction result for that entity pair to the relation in the self-owned knowledge. The reasons for using this method are: (1) intra-sentential reasoning is disturbed by contextual information from other sentences, so an independent model is needed to obtain intra-sentential self-owned knowledge; (2) integrating the acquired self-owned knowledge into the full-text relation extraction model to guide the relation extraction of other entity pairs is quite difficult for model construction; (3) if we replace the relation of an entity pair in full-text relation extraction when the probability is below a certain threshold, the problem is that the threshold is difficult to set; (4) we directly adopt the method of replacing with self-owned knowledge when the entity pair has no relation in the full-text relation extraction result. Although simple, this method has been verified to be effective through experiments. As shown in the right part of [Figure 3: see original paper], after full-text-level tokens undergo the reasoning-based academic full-text relation extraction steps described above, the final relation category of entity pairs is determined through the self-owned knowledge enhancement module.

The reasonableness of replacing entity pairs with no relation in full-text relation extraction results with relations from self-owned knowledge is based on two points: (1) compared with sentence-level relation extraction, full-text relation extraction is more challenging because it requires considering not only intra-

sentential logical relations between entities but also inter-sentential logical relations, involving long-range dependencies of entity pairs, which is more challenging [19,41]; (2) from existing public datasets, sentence-level experimental results are significantly 10%-20% higher than full-text-level (also called document-level) results [42], which to some extent indicates that sentence-level relation results have higher credibility than full-text-level relation results. Therefore, using simple entity relations extracted from the self-owned knowledge module to partially replace results extracted from the full-text-level module is reasonable.

4 Experiments and Analysis

4.1 Academic Full-Text Relation Extraction Datasets and Model Parameter Settings

Public datasets related to full-text relation extraction mainly include DocRED [43], CDR [44], and GDA [45]. All involve relation reasoning across multiple sentences with multiple entities, which is extremely challenging. DocRED is a large-scale dataset constructed from Wikipedia and Wikidata. CDR is a biomedical dataset constructed using PubMed, covering binary relations between chemicals and diseases, which is of great significance for biomedical research. The GDA dataset is also a binary relation classification task for identifying gene-disease interactions, constructed using distant supervision on MEDLINE, with lower quality compared to CDR. Our research focuses on relation extraction from academic full-text, so we selected the CDR and GDA datasets for experiments.

shows the data volume characteristics of the CDR and GDA datasets. The CDR dataset contains a total of 1,500 texts, equally divided into three parts: training set, validation set, and test set, all manually annotated. The GDA dataset contains a total of 30,192 texts. Compared with CDR, each text in GDA contains on average 2 fewer entities, 0.7 fewer entity mentions, while the average number of entity mentions per sentence is basically the same. It should be noted that we found many entity pair relations in the CDR test dataset were not annotated and were considered to have no relation. We re-examined these no-relation entity pairs and corrected 161 entity pair relations, calling this part of the test set CDR-revised.

We used Apex' s mixed-precision training method [46] for model training. lists some of the hyperparameters involved. All hyperparameters were tuned on the development set.

4.2 Comparative Experiments with Other Models

We compared our self-enhanced model ESOKRE with other similar research results, including BRAN [47], EoG [48], LSR [9], DHG [49], GLRE [50], SciBERTbase [51], and ATLOP-SciBERTbase [27]. As shown in (5 trials, taking the one with the highest F1 score), our self-owned knowledge enhancement model ESOKRE achieved an F1 score 11.13% and 0.35% higher than ATLOP-SciBERTbase on the CDR-revised and GDA datasets, respectively. The F1

score on the CDR dataset is relatively low because the dataset' s annotations are incomplete—when the self-enhancement model correctly identifies a relation, it is mistakenly considered as no relation because it is not annotated in the dataset. The improvement on the GDA dataset is not significant because the relations between genes and diseases in this dataset are relatively clear in the text, and simple reasoning can achieve good results, making it difficult for our self-owned knowledge enhancement capability to demonstrate its effectiveness. As shown in , the proportion of relations corrected by self-owned knowledge enhancement is 7.07% on both CDR and CDR-revised datasets. Although the number of corrected relations is the same for these two datasets, the F1 results differ because CDR-revised is a corrected version of CDR. On the GDA dataset, the proportion of relations corrected by self-owned knowledge enhancement is 0.40%, and the limited number of corrections affects the final F1 improvement.

4.3 Ablation Study

We conducted module ablation experiments to verify the effectiveness of different components of the proposed method. As shown in on the CDR-revised dataset, performance decreases when any module is missing. ESOKRE refers to automatically acquiring self-owned knowledge and enhancing full-text relation extraction through distant supervision. "ESOKRE - Self-owned knowledge enhancement" means using only full-text reasoning without integrating self-owned knowledge. "ESOKRE - SciBERT in self-owned knowledge"means using RoBERTa instead of SciBERT in the BERT module of the self-enhancement model. "ESOKRE - Graph module in self-owned knowledge" means removing the multi-view graph module from the self-enhancement model. "ESOKRE - Distant supervision" means removing the distant supervision module from the self-enhancement model. The results on the CDR-revised dataset show that the self-owned knowledge enhancement module and using SciBERT in the BERT module contribute the most to model performance. When they are removed from the entire model, the F1 score decreases by 6.72% and 4.32%, respectively. This indicates that our proposed self-owned knowledge enhancement module effectively assists full-text relational logic reasoning. Additionally, removing the distant supervision sub-module from the self-owned knowledge enhancement module decreases the F1 score by 1.91%, indicating that the distant supervision sub-module can improve the accuracy of self-owned knowledge acquisition to a certain extent.

It should be noted that in , adding the self-owned knowledge enhancement module to the CDR dataset actually decreases the F1 score. The reason, mentioned in Section 4.1, is that many entity pair relations in the CDR test dataset were not annotated and were considered no-relation, affecting the experimental results. Since our method shows limited improvement on the GDA dataset, this dataset is not experimented on in this and subsequent sections. The reason for the limited improvement has been mentioned in the previous section.

4.4 Analysis of Training Size Impact on Experimental Results

In the field of academic full-text knowledge graph construction, annotating training data is extremely time-consuming and labor-intensive. Therefore, it is necessary to evaluate the performance improvement of academic relation extraction models on limited annotated data. We compared the performance changes of ATLOP-SciBERTbase and ESOKRE with varying training text volumes. The experimental results are shown in . As the training text volume gradually increases from 10 to 500 pieces, the F1 scores of both models increase. When using the ATLOP-SciBERTbase method, the F1 score increases from 13.66 to 63.87. When using our ESOKRE model method, the F1 growth amplitude and rate are similar to ATLOP-SciBERTbase, specifically increasing from 15.83 to 75.83. The F1 data changes show that training data volume has a significant impact on academic text relation extraction, but continuously increasing training data volume does not necessarily lead to corresponding accuracy because the relationship between training volume and accuracy is not linearly increasing. This conclusion has great practical significance: in academic relation extraction work in a certain domain, it is necessary to evaluate the specific amount of manual annotation. This specific amount can be determined through an iterative approach—gradually annotating, training models, and evaluating performance. When the evaluation value reaches the inflection point of the growth curve, the final amount of data to be annotated can basically be determined.

4.5 Analysis of Self-Owned Knowledge Relation Replacement Methods

The research methods section mentioned that the method of using self-owned knowledge relations to replace candidate full-text relations was obtained through experimental verification. The experimental verification process is introduced below. We divided the methods of using self-owned knowledge relations to replace candidate full-text relations into three types: "replace when no relation," "replace when relation exists," and "direct replacement." These respectively mean: replacing when the candidate full-text entity pair result is no-relation but self-owned knowledge has a relation; replacing when the candidate full-text entity pair result has a relation but self-owned knowledge has no relation; and directly replacing the candidate full-text entity pair result with the self-owned knowledge relation. As shown in , on the CDR and CDR-revised datasets, replacing when no-relation is better than direct replacement, which is better than replacing when relation exists. The reason for this difference, we believe, is that the no-relation results obtained through long-range dependency logical reasoning of entity pairs can be supplemented by self-owned knowledge, while rashly replacing relations that exist in self-owned knowledge but not in long-range dependency reasoning is unreasonable because some relations are not reflected within sentences but between sentences.

4.6 Impact of Parameter Value in Distant Supervision Module on Relation Extraction Results

As mentioned above, in the distant supervision module, when of the top n search results simultaneously contain the entity pair, the relation corresponding to the entity pair is considered correct; otherwise, the entity pair relation is discarded. In experiments, top n is set to 10 and is set to 3. is set to 3 because this value yields the best F1 score. [Figure 5: see original paper] shows the comparison of different values in the distant supervision module on the CDR-revised dataset. The figure shows that when is 3, the final relation extraction F1 score is highest at 75.83%, while values less than or greater than 3 result in decreased F1 scores.

4.7 Case Study

We conducted case studies to further illustrate the effectiveness of our proposed ESOKRE model compared with the baseline model. As shown in [Figure 6: see original paper], both ATLOP (short for ATLOP-SciBERTbase) and ESOKRE can successfully extract the "drug-disease" relations between entity pairs "nafcillin" and "interstitial nephritis" and between "nafcillin" and "bacteremia." However, only our ESOKRE model can extract the "drug-disease" relation between entity pairs "daptomycin"and "bacteremia."This case shows that although the baseline model has certain reasoning capabilities on the CDR full-text entity relation matrix, it has deficiencies in specific cases. For example, the relation between "daptomycin" and "bacteremia" in this case can be inferred to be a "drug-disease" relation from sentence [3], although sentence [1] contains mentions of both entities but cannot clearly determine their relation. At this point, the context in sentence [1] interferes with the relation inference between the two. Our ESOKRE model first extracts self-owned knowledge, and through the operation of this module, the relation between "daptomycin" and "bacteremia" can be clearly determined.

As shown in the three comparison cases in , all listed are examples where the ATLOP model fails while our ESOKRE model succeeds. In the first case, entities "vancomycin" and "nephrotoxicity" clearly show a chemical-induced disease relation in sentences [1] and [12]. Similarly, in the second case, entities "acute renal failure" and "Chinese herbal" clearly show a chemical-induced disease relation in sentence [4]. In the third case, the relation between entities "cerebral vasospasm" and "cytarabine" can also be determined in sentence [4]. In summary, the reason why the ATLOP model fails is that it is affected by other sentences and entities in the context. With current long-text semantic analysis technology, it is difficult to effectively solve such problems with a single model alone.

5 Conclusion

This paper proposes an effective method for enhancing full-text relation extraction by using self-owned knowledge from academic full-text. The method obtains entity pair relations from sentences, places them in a search engine, uses distant supervision for further verification, and finally retains high-confidence entity pair relations as self-owned knowledge. Then, through reasoning modeling, adaptive threshold selection, and self-owned knowledge enhancement, full-text-level relation extraction is completed. Experimental results show that our proposed model achieves better performance than most existing models on the CDR-revised dataset. To our knowledge, this is the first attempt to integrate self-owned knowledge into academic full-text relation extraction. Future research plans mainly include: (1) studying linguistic logical reasoning to assist relational reasoning and further introducing dual-system theory from cognitive science to improve academic full-text relation extraction; (2) studying and comparing the differences between self-owned knowledge enhancement and external knowledge enhancement in academic full-text relation extraction; (3) adding performance comparison analysis of the model under multiple datasets from different disciplines.

References

[1] NUNOKAWA T, AKAZAWA M, YOKOGAWA N, et al. Late-onset scleroderma renal crisis induced by tacrolimus and prednisolone: a case report[J]. American journal of therapeutics, 2014, 21(5): e130-e133.

[2] ZHOU G D, SU J, ZHANG J, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the 43rd annual meeting of the association for computational linguistics (acl' 05). Michigan: ACL, 2005: 427-434.

[3] LI DONGMEI, ZHANG YANG, LI DONGYUAN, et al. A Survey of Entity Relation Extraction Methods[J]. Journal of Computer Research and Development, 2020, 57(7): 25.

[4] WANG JIANING, HE YI, ZHU RENYU, et al. Relation Extraction Technology Based on Distant Supervision[J]. Journal of East China Normal University (Natural Science Edition), 2020, 213(5): 122-139.

[5] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/1810.04805.pdf.

[6] QUIRK C, POON H. Distant supervision for relation extraction beyond the sentence boundary[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/1609.04873.

[7] PENG N, POON H, QUIRK C, et al. Cross-sentence n-ary relation extraction with graph lstms[J]. Transactions of the Association for Computational Linguistics, 2017, 5(1): 101-115.

[8] VERGA P, STRUBELL E, MCCALLUM A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction[EB/OL]. [2022-01-09]. https://arxiv.org/pdf/1802.10569.

[9] NAN G, GUO Z, SEKULIC I, et al. Reasoning with latent structure refinement for document-level relation extraction[EB/OL]. [2022-01-09]. https://arxiv.org/pdf/2005.06312.

[10] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach[EB/OL]. [2022-01-09]. https://arxiv.org/pdf/1907.11692.pdf.

[11] BELTAGY I, LO K, COHAN A. Scibert: a pretrained language model for scientific text[EB/OL]. [2022-01-10]. https://arxiv.org/pdf/1903.10676.

[12] EVANS J S B T, FRANKISH K E. In two minds: dual processes and beyond[M]. Oxford: Oxford University Press, 2009.

[13] XUE LU, SONG WEI. A Dynamic Label-Based Relation Extraction Method[J]. Computer Applications, 2020, 40(6): 1601-1606.

[14] SUN CHANGZHI. Joint Entity and Relation Extraction Based on Deep Learning[D]. Shanghai: East China Normal University, 2020.

[15] LIN Y, SHEN S, LIU Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th annual meeting of the association for computational linguistics. Berlin: ACL, 2016: 2124-2135.

[16] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th international conference on computational linguistics. Dublin: Dublin City University and Association for Computational Linguistics, 2014: 2335-2344.

[17] WANG L, CAO Z, DEMELO G, et al. Relation classification via multi-level attention cnns[C]//Proceedings of the 54th annual meeting of the association for computational linguistics. Berlin: ACL, 2016: 1298-1307.

[18] SOARES L B, FITZGERALD N, LING J, et al. Matching the blanks: distributional similarity for relation learning[EB/OL]. [2022-03-10]. https://arxiv.org/pdf/1906.03158.

[19] YAO Y, YE D, LI P, et al. DocRED: a large-scale document-level relation extraction dataset[EB/OL]. [2022-01-10]. https://arxiv.org/pdf/1906.06127.

[20] GUPTA P, RAJARAMAN S, SCHÜTZE H, et al. Neural relation extraction within and across sentence boundaries[C]//Proceedings of the AAAI conference on artificial intelligence. Hawaii: AAAI, 2019, 33(1): 6513-6520.

[21] XU W, CHEN K, ZHAO T. Discriminative reasoning for document-level relation extraction[EB/OL]. [2022-01-10]. https://arxiv.org/pdf/2106.01562.

[22] ZHOU H, XU Y, YAO W, et al. Global context-enhanced graph convolutional networks for document-level relation extraction[C]//Proceedings of the

57th annual meeting of the association for computational linguistics. Florence: ACL, 2019: 2649-2659.

[23] CHRISTOPOULOU F, MIWA M, ANANIADOU S. Connecting the dots: document-level neural relation extraction with edge-oriented graphs[EB/OL]. [2022-01-10]. https://arxiv.org/pdf/1909.00228.

[24] ZENG S, XU R, CHANG B, et al. Double graph based reasoning for document-level relation extraction[EB/OL]. [2022-01-10]. https://arxiv.org/pdf/2009.13752.

[25] GUO Z, ZHANG Y, LU W. Attention guided graph convolutional networks for relation extraction[EB/OL]. [2022-01-10]. https://arxiv.org/pdf/1906.07510.

[26] ZHOU W, HUANG K, MA T, et al. Document-level relation extraction with adaptive thresholding and localized context pooling[EB/OL]. [2022-03-01]. https://www.aaai.org/AAAI21Papers/AAAI-8308.ZhouW.pdf.

[27] YANG A, WANG Q, LIU J, et al. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. Florence: ACL, 2019: 2346-2357.

[28] WANG C, JIANG H. Explicit utilization of general knowledge in machine reading comprehension[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/1809.03449.

[29] WANG GUANYING. Relation Extraction with Knowledge Graph Embedding[D]. Hangzhou: Zhejiang University, 2019.

[30] CHEN Q, ZHU X, LING Z H, et al. Neural natural language inference models enhanced with external knowledge[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/1711.04289.

[31] WANG Z, LI L, ZENG D, et al. Knowledge-enhanced natural language inference based on knowledge graphs[C]//Proceedings of the 28th international conference on computational linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 6498-6508.

[32] PETERS M E, NEUMANN M, LOGAN I V R L, et al. Knowledge enhanced contextual word representations[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/1909.04164.

[33] YAO Y, YE D, LI P, et al. DocRED: A large-scale document-level relation extraction dataset[EB/OL]. [2022-03-01]. https://www.researchgate.net/profile/Zhenghao-Liu/publication/333815327_{DocRED}A_{Large}-Scale_{Document}-Level_{{{Relation}}{{Extraction}}}{Dataset}/links/5fc60274299bf1a422c77e3d/DocRED-A-Large-Scale-Document-Level-Relation-Extraction-Dataset.pdf.

[34] BELTAGY I, LO K, COHAN A. SciBERT: A pretrained language model for scientific text[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019: 3615-3620.

[35] XUE F, SUN A, ZHANG H, et al. GDPNet: refining latent multi-view graph for relation extraction[C]//Thirty-Fifth AAAI Conference on Artificial Intelligence. Online: AAAI, 2021: 2-9.

[36] LEE J, LEE I, KANG J. Self-attention graph pooling[C]//International conference on machine learning. Long Beach: ICML, 2019: 3734-3743.

[37] BAI LONG, JIN XIAOLONG, XI PENGBI, et al. A Survey of Distant Supervision for Relation Extraction[J]. Journal of Chinese Information Processing, 2019, 33(10): 10-17.

[38] MA JIN, YANG YIFAN, CHEN WENLIANG. Research on Person Attribute Extraction Based on Distant Supervision[J]. Journal of Chinese Information Processing, 2020, 34(6): 64-72.

[39] CHEN YUHENG, WANG ZHENG. Distant Supervision Relation Extraction Combining Attention Mechanism and Residual Networks[J]. Computer and Digital Engineering, 2020, 48(4): 909-913.

[40] GAO YONG. Research and Improvement of Enterprise Entity Relation Extraction Algorithm Based on Distant Supervision[D]. Shanghai: Shanghai Institute of Computing Technology, 2020.

[41] ZHANG N, CHEN X, XIE X, et al. Document-level relation extraction as semantic segmentation[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/2106.03618.pdf?ref=https://githubhelp.com

[42] Paperswithcode. Relation extraction | paperswithcode[EB/OL]. [2022-01-14]. https://paperswithcode.com/task/relation-extraction#datasets.

[43] YAO Y, YE D, LI P, et al. DocRED: a large-scale document-level relation extraction dataset[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/1906.06127.

[44] LI J, SUN Y, JOHNSON R J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction[J]. Database, 2016(1): 1-10.

[45] WU Y, LUO R, LEUNG H C M, et al. Renet: a deep learning approach for extracting gene-disease associations from literature[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/2009.10359.

[46] MICIKEVICIUS P, NARANG S, ALBEN J, et al. Mixed precision training[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/1710.03740.pdf?ref=https://githubhelp.com.

[47] VERGA P, STRUBELL E, MCCALLUM A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/1802.10569.

[48] CHRISTOPOULOU F, MIWA M, ANANIADOU S. Connecting the dots: document-level neural relation extraction with edge-oriented graphs[EB/OL]. [2022-03-01]. https://aclanthology.org/D19-1498/.

[49] ZHANG Z, YU B, SHU X, et al. Document-level relation extraction with dual-tier heterogeneous graph[C]//Proceedings of the 28th international con-

ference on computational linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 1630-1641.

[50] WANG D, HU W, CAO E, et al. Global-to-local neural networks for document-level relation extraction[EB/OL]. [2022-03-01]. https://arxiv.org/pdf/2009.10359.

[51] BELTAGY I, LO K, COHAN A. Scibert: a pretrained language model for scientific text[EB/OL]. [2022-03-01]. https://aclanthology.org/D19-1371.pdf.

Author Contributions:
Zhuo Keqiu: Proposed research ideas and methods, wrote initial draft;
Shen Si: Proposed paper ideas and revisions;
Wang Dongbo: Paper revision and review.

Note: Figure translations are in progress. See original paper for figures.