

## Annotation and Joint Extraction of Scientific Entities and Relationships in NSFC Project Texts

**Zhiyuan GE**

*School of Economics and Management, Beijing University of Technology, Beijing 100022, China*  
*E-mail: gezhy@bjut.edu.cn*

**Xiaoxi QI**

*School of Economics and Management, Beijing University of Technology, Beijing 100022, China*  
*E-mail: qxx81999@163.com*

**Fei WANG**

*School of Economics and Management, Beijing University of Technology, Beijing 100022, China*  
*E-mail: 1056756589@qq.com*

**Tingli LIU**

*School of Economics and Management, Beijing University of Technology, Beijing 100022, China*  
*E-mail: liutingli@bjut.edu.cn*

**Jun GUAN**

*School of Economics and Management, Beijing University of Technology, Beijing 100022, China*  
*E-mail: guanjun@bjut.edu.cn*

**Xiaohong HUANG**

*Office of Academic Affairs, Beijing University of Technology, Beijing 100022, China*  
*E-mail: huangxh@bjut.edu.cn*

**Yong SHAO**

*Faculty of Information Technology, Beijing University of Technology, Beijing 100022, China*  
*E-mail: mazel@bjut.edu.cn*

**Yingmin WU**

*School of Management, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China*  
*E-mail: amynii@qq.com*

**Abstract** Aiming at the lack of classification and good standard corpus in the task of joint entity and relationship extraction in the current Chinese academic field, this paper builds a dataset in management science that can be used for joint entity and relationship extraction, and establishes a deep learning model to extract entity and relationship information from scientific texts. With the definition of entity

---

Received March 15, 2023, accepted June 1, 2023

Supported by the National Natural Science Foundation of China (71804017), the R&D Program of Beijing Municipal Education Commission (KZ202210005013), and the Sichuan Social Science Planning Project (SC22B151)

and relation classification, we build a Chinese scientific text corpus dataset based on the abstract texts of projects funded by the National Natural Science Foundation of China (NSFC) in 2018–2019. By combining the word2vec features with the clue word feature which is a kind of special style in scientific documents, we establish a joint entity relationship extraction model based on the BiLSTM-CNN-CRF model for scientific information extraction. The dataset we constructed contains 13060 entities (not duplicated) and 9728 entity relation labels. In terms of entity prediction effect, the accuracy rate of the constructed model reaches 69.15%, the recall rate reaches 61.03%, and the F1 value reaches 64.83%. In terms of relationship prediction effect, the accuracy rate is higher than that of entity prediction, which reflects the effectiveness of the input mixed features and the integration of local features with CNN layer in the model.

**Keywords** joint extraction of entities and relations; deep learning; Chinese scientific information extraction

## 1 Introduction

Artificial intelligence has promoted the intersection of disciplines with the development of machine learning and deep learning, especially in discipline synthesis and method application, and it helps in analyzing the relationships between different fields of scientific research and promotes discipline integration<sup>[1]</sup>. Scientific literature, especially academic documents such as papers and project proposals, contains rich information in scientific fields. If such information can be fully mined, it will help scientists find relevant papers and fields and automatically understand the key ideas of relevant research. NLP (Natural Language Processing) is an effective tool in these activities, and information extraction technology in NLP has become one of the main ways to transform the unstructured information of scientific papers into structured information.

The effective excavation of a large amount of tacit knowledge in scientific literature can enhance the academic researchers' scientific research capability. With the rapid development of science and technology, the growth rate of scientific publications is also accelerating<sup>[2]</sup>. It is a great challenge for researchers to improve the reading quality and speed of scientific literature, and to find useful information such as methods and ideas from existing research for their own studies. Therefore, scholars have started to explore the information in scientific texts, such as titles, sentences, abstracts, and even the full paper text of papers, and have also experienced a long process of exploration. Existing studies on knowledge extraction in academic literature focus on the identification of topics, keywords and terms<sup>[3]</sup>, or the co-occurrence analysis by word frequency statistics, topic clustering, chi-square test, and other statistical methods<sup>[4]</sup>. In the studies of Chinese academic texts, the relationship between entities, such as terms and topics, is also limited to the study of co-occurrence and hierarchical relationship, which limits the ability of knowledge discovery between entities<sup>[5]</sup>. In addition, these studies have provided a relatively coarse grain level in entity recognition<sup>[3]</sup>.

Although some scholars focus on named entity recognition (NER) in the academic field, there are also two main problems compared with general named entity recognition. The first is the lack of a relevant large corpus<sup>[1]</sup>. For example, the English Annotated ScienceIE dataset consists of only 500 scientific paragraphs and contains only three entity categories<sup>[6]</sup>, and these corpus datasets do not explain their classification criteria. The second problem is inconsistent

labeling. The inconsistency of labeling is mainly reflected in three aspects, the first one is the inconsistency of classification of labeling<sup>[7]</sup>, that is, the same word is labeled differently in different corpora, for example, an entity labeled as “technology” in the ACL dataset is labeled as “technology and method” in the Semeval 2018 dataset; The second one is that different people may have different views on labeling the same entity, for example, in the corpus mentioned by Luan, entities are classified as task, process, material, and Luan says that one person might think of it as a task, another person might think of it as a process<sup>[6]</sup>. The form of labeling is also inconsistent. Some people use BIO labels and others use BIOES labels.

Named entity recognition is also very different in different languages<sup>[8]</sup>. Compared to English, Chinese has many different features that make it difficult for NER<sup>[7]</sup>. For example, word segmentation is a natural drawback in Chinese<sup>[6]</sup>, and as a result, the length of the entities is variable, and the academic entities may be longer. In addition, the number of named entities in the academic field is usually not small, and the number of academic journals published each year is large and growing. However, the current Chinese academic corpus is usually small and basically not open.

In conclusion, the current research on natural language processing in the academic field is relatively scarce, and the results of studies on entity recognition and relationship extraction of scientific documents have not yet supported NER application tasks, especially in Chinese scientific documents. Our study uses the basic information of projects funded by the National Natural Science Foundation of China as the main data source to build a corpus of entity relationships, including titles, keywords, and abstract texts. And we continue the studies on the joint extraction of entity relationships in Chinese academic fields in management science. In this paper, we first classify the entity and relationship categories of academic texts, and then build the corpus dataset based on manual annotation. Then, we build the end-to-end sequence annotation model based on BiLSTM-CRF, and perform the joint extraction of entity relations with our academic text corpus. To improve the effectiveness of our model, we add a new label named entity clue word vector during the corpus construction in the process of manual annotation, which strengthens the relationship extraction and solves part of the overlapping relationship problem with this new annotation method.

## 2 Related Work

### 2.1 Entity Recognition and Relation Extraction

The basic methods of NER have involved from rule-based, dictionary matching, machine learning to deep learning<sup>[9]</sup>, or other categories including rule-based, statistical learning based, hybrid, and deep learning based<sup>[10]</sup>. Research on related aspects in entity recognition and relation extraction are generally divided into two tasks and be studied independently.

The study of entity recognition has undergone a shift from rule-based, machine-learning-based to deep-learning-based. Current deep learning based methods have become mainstream methods in tasks such as NER. Here are some deep learning methods that have been used in information extraction during recent years: CNN-BiLSTM-CRF<sup>[11]</sup>, LSTM-CRF<sup>[12]</sup>, BiLSTM-CRF<sup>[11, 13, 14]</sup>, BiGRU-CRF<sup>[15]</sup>. The above methods all require a manually annotated corpus. There is another method that does not use manual annotation, which uses classification methods

in machine-learning to directly perform text extraction. For example, Lee, et al. used an iterative method to extract Chinese entities from texts, unlike traditional entity extraction methods which require a dictionary and extraction rules, they use support vector machine classifiers to train extraction rules from terms that appear in the corpus<sup>[16]</sup>. Geng, et al. used a combination of semantic-based and structure-based algorithms to extract cross-domain ontologies<sup>[17]</sup>. Nasar, et al. performed the information extraction based on the metadata and full-text of ACM and IEEE academic articles (scientific articles), where the metadata is created from the semi-structured format of scientific articles, and the key-insights extraction is established from the full-text<sup>[18]</sup>.

The joint learning method combines the two tasks of entities and relations into a single model, which can reduce error propagation and can capture the correlation between entities and relations by sharing the bottom features<sup>[19, 20]</sup>. These methods are not only applied to English texts, but also widely used in the joint extraction of entity relations in Chinese academic texts. For example, Zhang, et al. combined convolutional neural networks with support vector machines and conditional random fields to build models in the parameter sharing mode, and conducted research on the joint extraction of entities and relations in the field of Chinese medicine, and achieved good results in experiments on the drug specification corpus, and the F1 scores of entity classification and relation extraction reached 98.0% and 98.3%, respectively<sup>[21]</sup>. Liu, et al. constructed a Chinese medical text joint extraction model based on the label scheme, which complemented the joint extraction study in the field of Chinese medicine with different research methods<sup>[22]</sup>. Zhou, et al. proposed a joint model BERT-LCM-Tea for entity and relation extraction to solve the problems of multi-meaning of words and relation overlap in Chinese tea texts, and the F1 value achieved good results, scoring 86.8% in the entity recognition task and 77.1% in the relation extraction task<sup>[23]</sup>.

## 2.2 Scientific Corpus Dataset

The categories of entities in scientific text corpus have different standards. Zadeh, et al. proposed a scheme of seven entity categories as methods, tools, data sources, datasets, models, evaluation metrics, and other types<sup>[24]</sup>. The entity types of ScienceIE dataset are Task, Process, Material<sup>[6]</sup>, and in Semeval 2018 dataset there are seven types of entities as technology and method, tool and library, language resource, language resource product, measures and measurements, models and other. In the SciERC corpus dataset, the entities were divided into 6 categories as task, method, evaluation metric, material, other scientific terms, and generic terms. Nasar, et al. listed some previous scholarly classifications of academic corpus, including Objective, Method, Result, Conclusion, etc., which are basically consistent with the structure of academic papers<sup>[18]</sup>. And the dataset ACL Anthology Reference corpus has 6 types of entities: Technology, system, language resources (specific product), model, measurement and other<sup>[25]</sup>.

There is a lot of research on the categories of relations between entities in English corpus. Kruiper, et al. divide the relations in scientific information datasets into three categories: Trade-off, Argument-Modifier, Not-a-Trade-off<sup>[26]</sup>, which are based on FOBIE, SciERC, ScienceIE, and Semeval 2018 datasets. In the Semeval 2018 dataset, the given entity relations are divided into 6 types: Usage, result, model, part-whole, topic, and comparison. In the SciERC dataset, 7 relations are defined: Used-for, feature-of, hyponym-of, part-of, compare, conjunction and

coreference.

Some scholars have proposed other types of divisions for the entities in Chinese scientific corpus. For instance, Zhao, et al. extracted only theory terms from Chinese Wikipedia, journal articles, and dissertations<sup>[27]</sup>, without specifying the origin of these journal articles and dissertations. Zhang, et al. extracted only research method entities<sup>[28]</sup>, and their corpus included 198 journal articles from the Journal of Information Technology. In another article Zhang classified entities into the following four categories: Method entities, tool entities, resource entities, and indicator entities<sup>[29]</sup>. Jiang, et al. used 25 articles from the Journal of Information Science to build corpus and extract entities and relations according to four entity types: Methods, tasks, tools, and resources<sup>[5, 30]</sup>.

In fact, recent studies<sup>[31, 32]</sup> provide an in-depth analysis and discussion of the methods, corpus, and other aspects of entity relationship extraction. They do this by providing a more thorough review of the literature on entity relationship extraction in scientific texts in recent years. It is recommended to read [31, 32] for more information.

In general, from the perspective of annotation, the basic standards of categories of entities and relations still have not yet reached consistency in different languages, and especially there is no research on the labeling norms in Chinese scientific texts, and researchers in extracting information from Chinese scientific texts also do not adopt the classification method from English corpus dataset. From the perspective of corpus dataset for scientific texts, there are some foundations for English scientific information extraction, and some datasets have been built in this area, including ScienceIE, SemEval, SciERC, FOBIE, etc.<sup>[6, 25, 26]</sup>. Zadeh also listed some resources available for term extraction and named entity recognition<sup>[24]</sup>. Table 1 lists brief information about some corpus datasets.

Actually, we have not found any open dataset of Chinese scientific corpus that can be used for scientific information extraction.

### 3 Materials and Methods

#### 3.1 Methodology

For the information extraction task of entities and relations in Chinese scientific documents, we need to address two fundamental issues: Where can we find the corpus to be used for learning and training? What is the model made of?

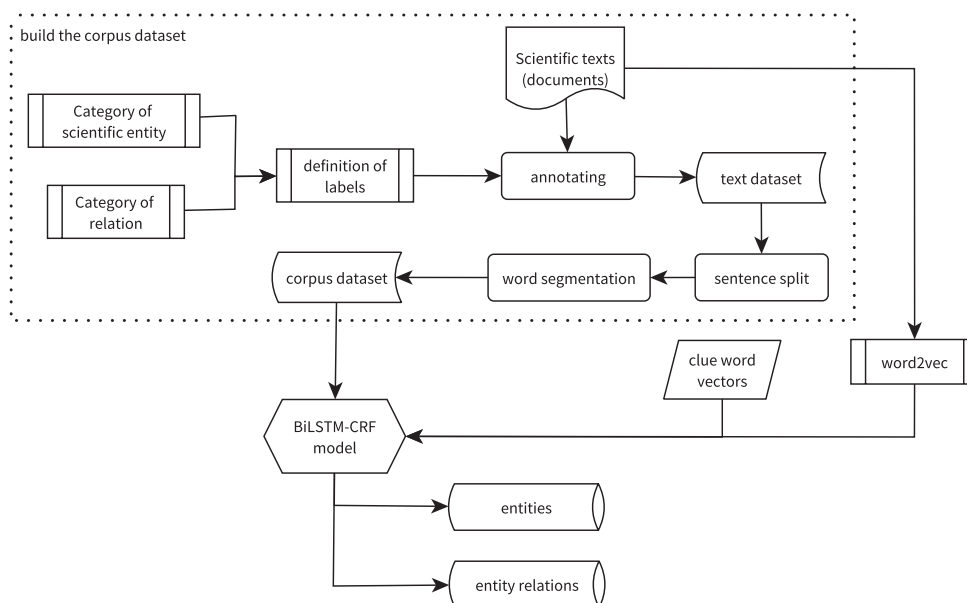
In the previous discussion, we found that there is no public Chinese dataset for the current corpus in the scientific field, and some scholars in this field use self-built corpus dataset, which is not open for free use. In addition, in the process of self-building corpus, there is no unified standard or approved annotation scheme for labeling Chinese scientific corpus, and previous researches in China do not adopt or refer to the annotation classification method in the English corpus. Therefore, the researches on entity relationship extraction between Chinese and English cannot be compared and analyzed.

This study initially addresses the issue of label specifications and how to classify corpus annotation labels using the results of recent investigations. Then, to perform the joint extraction of entities and relations for Chinese scientific documents, we build a professional corpus and design a joint extraction model of entity relations based on deep learning techniques. The

research idea of this paper is shown in Figure 1.

**Table 1** Entities and relations categories in some scientific corpus dataset

Dataset	Entity types	Relation types
ScienceIE ( <a href="http://scienceie.github.io/">http://scienceie.github.io/</a> )	Task, process, material	-
ACL ( <a href="https://aclanthology.org/">https://aclanthology.org/</a> )	technology, system, language resources (specific product), model, measurement and other	-
SciERC ( <a href="https://github.com/allenai/scierc">https://github.com/allenai/scierc</a> )	task, method, evaluation metric, material, other scientific terms, and generic terms	used-for, feature-of, hyponym-of, part-of, compare, conjunction and coreference
Semeval 2018 ( <a href="http://scienceie.github.io/">http://scienceie.github.io/</a> )	technology and method, tool and library, language resource, language resource product, measures and measurements, models and other	usage, result, model, part-whole, topic and comparison
PubMed ( <a href="https://pubmed.ncbi.nlm.nih.gov/">https://pubmed.ncbi.nlm.nih.gov/</a> )	Chemicals, diseases, anatomy, procedures, organisms, companies, locations, time, numeric values	Chemical-disease, Gene-disease, Anatomical part-located in, Process-performs on, Mutual interactions, Temporal relations, Causal relations, Comparative relations
Jiang (Chinese)	Method, task, tool, resource	-
Zhang (Chinese)	Method, tool, resource, measurement	-



**Figure 1** Framework of joint extraction of entity and relation for Chinese scientific texts

The first part of our research in Figure 1 is to construct the corpus for extracting Chinese entities and relations in the scientific documents. At first, we discuss the basic elements of scientific articles such as research objects and methods, and define the categories of entities in scientific documents. According to the matching rules between the elements of the documents, the relationship categories between the entities are given. And then we propose an annotation scheme to manually annotate the Chinese scientific texts to build the corpus dataset.

In the second part in Figure 1, we construct a deep learning model of BiLSTM-CRF, in which the input layer integrates character-based vector based on word2vec and information of clue word vectors that are special features in Chinese, and then introduce a convolutional neural network (CNN) layer in the BiLSTM-CRF model to strengthen the model's extraction of local word features, thereby improving the effect of joint extraction of entity relations.

The final output of the model is the corresponding label in each sentence in the text, and the result of extracting these labels into entities and relations needs to be matched. While the downstream tasks after entity and relation extraction in NER require the triple form which includes entities and their relations like (entity1, relation, entity2), we also define corresponding matching rules to get a triplet output of the predicted results.

## 3.2 Annotation Categories of Scientific Entities and Relations

### 3.2.1 Data resource

From the perspective of academic science, the projects of the National Natural Science Foundation of China can represent some of the current academic research frontiers and research levels, and this information is public on the official website of NSFC. Therefore, we collect the data of project titles, keywords and abstract texts in management science in 2018 and 2019 from the NSFC website to build our initial corpus text data.

To test and supplement our dataset, we also collect some abstract texts from the Chinese journal Library and Information Work. Our dataset currently contains a total of 1076 articles. In future research we are considering adding the project data of NSFC which is not currently included, and building a more complete dataset for scientific terms in the field of management science.

### 3.2.2 Definitions and categories of Chinese scientific entity and relation

In the definition of entity types, we have listed some datasets in Table 1, including ACL, ScienceIE, Semeval 2018, SciERC, etc., and most of them refer to the basic elements or structure of scientific articles, i.e., such as purpose/objective, methodology, data resource. Kruiper also introduced some annotations of English scientific datasets<sup>[26]</sup>. Referring to these studies in entity types, we believe that the basic elements are materials, processes and tasks<sup>[6]</sup> as dataset ScienceIE. In all these types of entities, the most basic elements are related to tasks and methods, because every paper or research project aims to solve a problem or achieve a certain goal, or to accomplish a task, and to do this, some theories, methods, tools, resources, models, etc. are used, which are commonly known as methods and tools. Therefore, in this paper we will discuss two types of entities: Research objects, methods and tools. Here objects refer to tasks, or these things research problems discussed, and methods and tools refer to things used to

accomplish tasks or solve problems.

The research object in a task or problem in a scientific article or project is defined as the part, several parts, or the whole of the specific thing to be studied. This paper classifies the research objects into three types of entities: Dimensional entity, subject entity, and purpose entity. While the article or project discusses some factors or the subject can be divided into subsets, the dimensional entity is defined as the subset or superset of things. The subject entity is defined as the main research object of the article. And the purpose of the research wants to achieve and the results obtained are considered as the purpose entity.

The definition of the entity of methods and tools in the previous researches can be divided into two types: The first type is methods, tools or techniques to solve problems in the application field; the other type is the ideas or solutions to the problem proposed by the author. Based on the first definition, we classify the entity of methods and tools into theoretical entity, basic method entity and usage entity. Therefore, the final entity types classified in this paper are divided into six types, including research subjects, purpose entities, dimensional entities, base method entities, usage entities, and base theory entities. The specific information of the six academic domain entity types in the self-constructed corpus of this paper is shown in Table 2.

**Table 2** Types of entities in scientific texts

Type of Entity	Description of the entity
Research subject	Using ... as a research object, research ... issues
Purpose entity	Obtained entities or the purpose of the study
Dimensional entity	An entity representing a part in a local-whole relationship
Basic method	Based on the ... (the basic method) of ... methods
Usage entity	Based on the ... of ... (method of use) method
Basic theory	Game Theory, The theory of ...

We also refer to the classification definitions of entity relations in these datasets such as Semeval 2018, SciERC, etc. Semeval 2018 divides relations into 6 categories: Usage, result, model, part-whole, topic, comparison, SciERC divides relations into 6 categories: Used-for, feature-of, hyponym-of, part-of, compare, conjunction, coreference. Actually, all these relation categories should be based on the entity categories. Contrary to the categories of this dataset, we had first considered the relations of each pair between all entity types.

The type of relationship is determined based on the type of entity and the relationships between the pairs of entities. For the entities of the research object entities, we can form most of the relationship pairs: The research subject entity can be paired with the dimensional entity, the research subject itself, the purpose entity, the basic method/usage method, and the research purpose entity can be paired with the research method, the basic theory entity, and so on. Among the research method entities, the basic method entities and the usage method entities can be paired to define a relationship.

The types of relations between the entities we consider in this paper are described below. 1) When a paper expresses an entity pair with holistic and part-of relation types, the entity of the part-of relation is regarded as a dimensional entity and the entity of the holistic relation is the



research subject entity, and the entity pair is defined as having a dimensional relation. 2) When there is a contrastive, comparative, and associative type of relationship between two research subjects, the combination of the entity pair is defined as having an associative relationship. 3) When there is a causal type of relationship between the research method or subject and the purpose of the research, the pair of entities of this type is considered to have a purposive relationship. 4) When a type of research method is applied to a research subject, we consider that there is a using relationship between the pair of entities. 5) When a basic method is improved so that the final method entity of the article is obtained, we consider that there is an improving relation between the basic method and the using method. 6) When a certain theory or knowledge is used to achieve a certain research purpose or subject research, we consider that there is a supporting relationship between the pair of entities of this type. In addition, in order to solve the overlapping problem of entity one-to-many and many-to-one, we introduce the overlapping relationship type. 7) When an entity has a relationship with multiple entities, the entity is considered to have an overlapping relationship. Other types of relationships are not considered in this paper.

Finally we defined 7 types of entity-pair relationships, and the specific explanation of each relationship type is shown in Table 3.

**Table 3** Types and descriptions of entity relations

Type of relation	Entity pairs	Common clue words	Description
Dimensional relation (WD)	dimensional entity-research subject	from, dimension	There is a part-of relation between the research objects
Associated relation (GL)	research subject-research subject	The relation/ influence between ...	There is a direct relationship between research subjects
Purpose Relation (MD)	research method/subject - purpose entity	get, build	Use a tool or perform an operation for a specific purpose
Usage relation (SY)	research method - research subject	apply, use	Use a method to study an subject
Improved relation (GJ)	basic method - method used	refer to, on the basis of ...	Improving on some methods to get the final method used
Relation of support (ZC)	theory - purpose	analysis, prediction	Use a theory to support a result
Overlapping relation (OVE)	entities have multiple relations with other entities		The entity has multiple relations with other entities and exists in multiple triples

Our classification definition has the following advantages: 1) all categories of the entity are clearly defined, so that labeling is less controversial; 2) our definitions have taken into account the relationship of entities, so that the following definitions of categories of entity relationship can be simplified, making manual labeling become easy and clear.

### 3.2.3 Annotation of scientific corpus

In order to adapt to the characteristics of dense entity distribution and diverse entity types in the academic field, we use the BIO annotation method, which is the most widely used in the field of entity recognition. The location annotation of the entity, that is, the beginning of the entity and the internal word position. If a B (Begin) appears continuously with one or more I (Inside), it is able to find the location range of a complete entity. B (Begin) indicates the start position of the entity to be extracted, I (Inside) indicates the internal position of the entity to be extracted, and O (Other) indicates that the token under this tag does not belong to the entity to be extracted.

The relation information is represented using the types predefined over in this section, as detailed in Table 3. If “relation information + role information” in the label is “OVE-1”, it means that the entity with the label is in different pairs of entities, and acts as the main role in one pair of relations and the object role in another pair; if “relation information + role information” in the tag is “OVE-2” in the tag, it means that the entity under the tag has played the same role in the process of combining entities in the team, that is, the object role, and has been in the state of action reception.

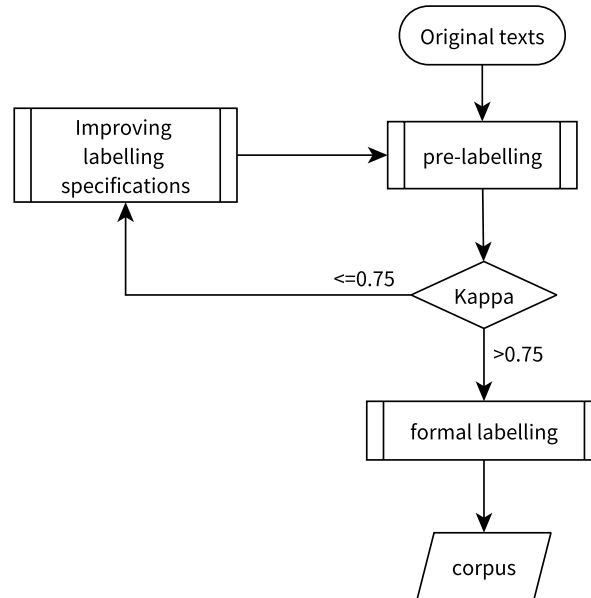
The role information conveys the subject and object of the entity in a triple, that is,  $\langle \text{entity 1, relationship type, entity 2} \rangle$ , entity 1 is the subject role in the triple, entity 2 is the object role in the triple, and the role information can be used to distinguish the type of entity and to convey the action state of the entity to some extent. Sending and receiving of an action is a corresponding set of states, so it has to appear in pairs. There are 29 types of actual combination labels, as shown in Table 4.

**Table 4** Type of annotation labels

Label types	Annotation labels	Number of types
Non-entity tag	O	1
General entity labels	B-GL-1, I-GL-1, B-GL-2, I-GL-2, B-GJ-1, I-GJ-1, B-GJ-2, I-GJ-2, B-MD-1, I-MD-1, B-MD-2, I-MD-2, B-WD-1, I-WD-1, B-WD-2, I-WD-2, B-ZC-1, I-ZC-1, B-ZC-2, I-ZC-2, B-SY-1, I-SY-1, B-SY-2, I-SY-2	24
Overlapping relationship labels	B-OVE-1, I-OVE-1, B-OVE-2, I-OVE-2	4

In the process of annotation type definition and corpus annotation, our annotation specification verification process is shown in Figure 2. In this process, three members of the paper team selected nearly 160 pieces of data for trial annotation, and used the Kappa index to test the consistency of the annotated data set to test the usability of the annotation specification for the consistency of the annotation results. The annotators first discuss the annotation specifications originally defined, and then each annotator independently annotates, through the consistency calculation of the 160 pieces of data of the annotation results into entity units, if the consistency is found to be unable to pass the requirements, then discuss the specification definition of the annotation again, improve and determine the specifications, until the final

annotation specification is formed. In the annotation consistency statistics in this paper, the goal is mainly to see whether the statistical relationship type is consistent, and the relevant information of entity boundary determination is not reflected in the statistical result set, but the division specification of entity boundary and entity type has been carried out through the improvement of the specification.



**Figure 2** Inspection process of marking specification

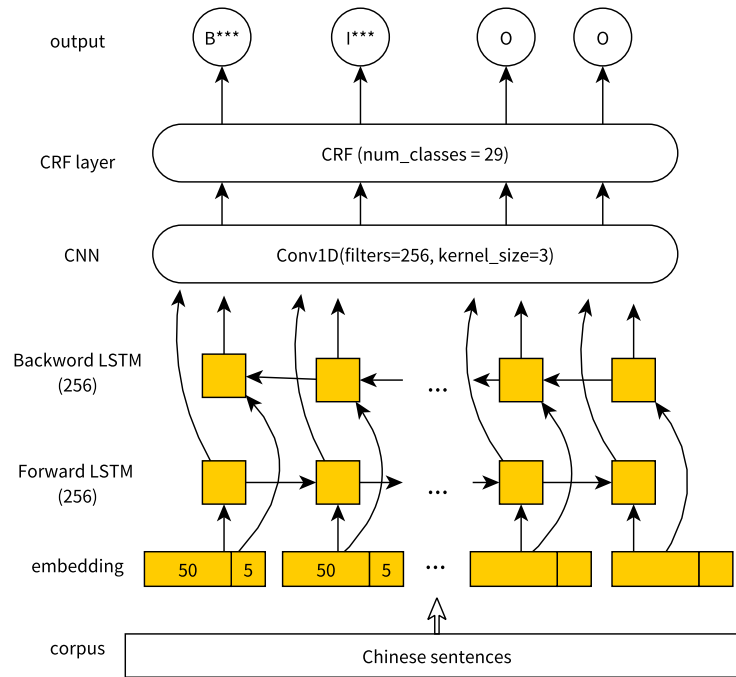
The formula for Kappa index is as follows:

$$K = \frac{P(A) - P(E)}{1 - P(E)}, \quad (1)$$

where  $P(A)$  represents the actual observation of the consistency of the labeling results, and  $P(E)$  represents the expected value of the consistency of the labeling results. If  $K \leq (q) 0.75$ , the labeling result is reliable, and if  $K \leq (q) 0.4$ , the labeling result is more reliable. For each dataset, through the annotation and discussion of entities and relationships three times, the consistency calculation of 160 pieces of data independently labeled by annotators is carried out on an entity-by-entity basis, and the calculation results show that the Kappa coefficient is 0.77, and the standardization of the labeling results is reliable, so that the standard specifications of the above-mentioned entities and relationships are finally formed, and large-scale data labeling can be carried out.

### 3.3 Model

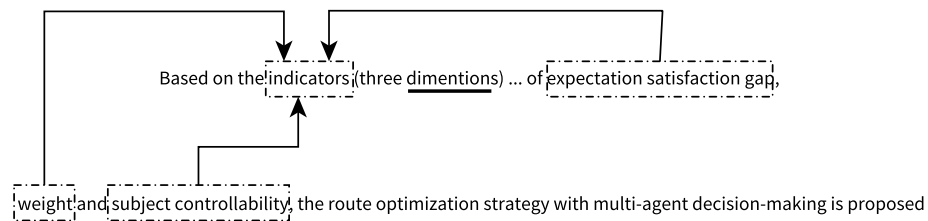
The structure of the model used in this paper contains four layers: Embedding layer, BiLSTM layer, CNN layer, and CRF layer, as shown in Figure 3. In our model, the embedding layer and BiLSTM layer are used to acquire for contextual features, and the CNN convolutional layer is used to enhance the local features to improve the model's representation of semantic information and features, and the CRF layer is used for the sequence annotation task.



**Figure 3** BiLSTM-CNN-CRF model

### 3.3.1 Embedding layer

In general, there are obvious writing specifications for texts in academic fields, and these specifications are very helpful for feature extraction. For example, when judging the relationship between pairs of entities, we find that there are certain words that can help us judge the existence of the relationship. In the example sentence shown in Figure 4, when the underlined word “dimension” is present, there is a high probability that a pair of entities with a dimensional (WD) relationship will appear nearby, and “expected satisfaction gap”, “weight”, and “subject” can all appear as dimensional entities, and there is a dimensional relationship with “indicator”. The words “construct”, “factor”, “research”, “influence”, and so on in the Chinese sentence are also like this, which are defined as clue words.



**Figure 4** Example of clue words features

Therefore, the scheme of this article not only uses common word vectors as feature inputs, but also considers such words that can help us determine the existence of relationships. Thus, the input features of our embedding layer are composed of word vectors plus clue word features.

The clue word features are manually extracted from our clue word database which obtained through word frequency statistics, and if a word is in our clue word database, we append a 5-bit feature filled with '1' to strengthen the information, otherwise we append a 5-bit feature filled with '0'.

For the features of the word vectors we choose to use the CBOW model of Word2Vec with Gensim to train the corpus. For the selection of the word vector dimensions, this paper refers to the conclusion of the principle of minimum entropy: The vector length should follow the formula  $n > 8.33 \log N$ , where  $N$  represents the size of the word list. The size of the word list used in this paper is 58314, so the dimension of the word vector is calculated as 39.70, and considering the general length chosen in NLP algorithms here we choose 50 as the vector length. In addition, to make sure that the dimension of the word vector does not have a bad effect on the model results, we have conducted a series of experiments on the word vector dimensions of 10, 30, 50, 100, 200 and 300, and it was found that the accuracy of the results are not much different, but the difference in the training time of the model is large. Therefore, the word vector with 50 dimensions was finally selected for the final model training.

The final word embedding form is a 55-dimensional vector formed by the splicing of a 50-dimensional word vector and a 5-dimensional manual clue word feature  $V_i = \langle V_W^i | V_C^i \rangle$  as the final distributed word vector representation in this paper.

### 3.3.2 BiLSTM layer

Long short-term memory network (LSTM) improves the disadvantage of recurrent neural networks (RNNs) in gradient explosion or gradient disappearance during computation by introducing the gates<sup>[33]</sup>. In this paper, forward LSTM is used to train sentences in positive direction to obtain the previous text features of words  $[\vec{h}]$ , while backward LSTM is used for training sentences backward direction to obtain the following text features of words  $[\overleftarrow{h}]$ . The previous text features and the following text features of the word are concatenated to obtain the contextual features of the word  $[\vec{h} | \overleftarrow{h}]$ .

### 3.3.3 Convolution layer

The advantage of convolutional neural networks lies in two features of the design of the network itself, weight sharing and local connection. Convolutional neural networks rely on local connections to significantly reduce complexity and perform automatic feature extraction. The network structure of convolutional neural network generally adopts convolutional layer, pooling layer and activation function layer to be alternately stacked with each other. This paper uses a convolution to perform convolution calculation of features, and connects the fully connected layer to construct the convolutional neural network layer.

Since the effective length of each sentence in texts is different, the position of entities may appear in a different place for each sentence, while the convolutional neural network layer helps to complement the local features at short distance, and performs well in entity and relationship extraction tasks.

### 3.3.4 CRF layer

Conditional random fields (CRF) are usually used as label decoders, the output of the previous convolutional layer is used as the input of this layer in our model, and the maximum score sequence of the sentence is used as the final label sequence for the output. For learning, the conditional probability model  $\hat{P}(Y|X)$  is obtained by using the training data through maximum likelihood estimation or regularized maximum likelihood estimation. For prediction, it is used to find the output  $y$  with the maximum conditional probability  $\hat{P}(y|x)$  for a given input sequence  $x$ .

## 4 Results and Analysis

### 4.1 Dataset

Our dataset uses the data annotation results according to the annotation labels in Subsection 3.2 of this paper as a dataset for the joint extraction of entity relations in the Chinese scientific texts. Our dataset contains 5,000 sentences, and six types of entities, a total of 13,060, and seven types of relations, a total of 9,728. Table 5 shows the details of entities and relations in our dataset.

**Table 5** Number of entities and relations in our dataset

Entity types	Quantity	Type of relations	Quantity
Research subject	6100	Dimensional relation (WD)	1198
Purpose entity	2204	Associated relation (GL)	2517
Dimensional entity	937	Purpose relation(MD)	1007
Basic method	569	Usage relation (SY)	2489
Usage method	2410	Improved relation (GJ)	900
Theory	840	Relation of support (ZC)	1102
–	–	Overlapping (OVE)	515
Total	13060	Total	9728

### 4.2 Measurements

Classical model evaluation tools commonly used in information extraction are accuracy ( $P$ ), recall ( $R$ ), and F1 value, which are shown as Equations (2), (3), (4) as follows. In these equations, TP represents the number of entities or labels that were correctly predicted in the prediction results, FP represents the number of entities or labels that were incorrectly predicted in the prediction results, and FN represents the number of entities or labels that were not recognized in the real results.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (3)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

In general, the measurements of accuracy and recall influence each other, and in order to consider the results of both aspects, we show the results of the three indicators, but the comparison of results focuses more on the F1 value.

### 4.3 Parameters

The experiments were implemented on the anaconda programming tool, using Python version 3.8 and TensorFlow version 2.2.3. The hyperparameter settings are shown in Table 6.

**Table 6** Experimental parameter settings

Hyperparameter	Value
Batch_size	8
Word2vec_size	50
Clue_size	5
BiLSTM_layer	256
Vocab_size	58314
max_len	232
epoch	15
Learning_rate	0.001
validation_split	0.25

Batch\_size is set to 8 according to the capacity of the CPU, Word2vec\_size is the word vector dimension 50 for Word2vec training, Vocab\_size size is the corresponding size of the lexicon, max\_len size is the maximum length 232 in all sentences, epoch is set to 15, the learning\_rate is set to 0.001 based on the experience of information extraction model.

### 4.4 Results

In order to verify the use of each feature in this study and the effects of our model, a series of experiments are set up to compare the results. For the comparison of entity and relation joint extraction models, we selected the BiGRU-CRF model, CNN-CRF model, and BiLSTM-LSTM model (softmax) to validate the experimental results. In Word2Vec and the validity of clue word features, we conducted comparison experiments on initialization embedding and using only word vectors as conditions, respectively. The results of each comparison experiment are shown in Table 7.

Table 7 shows the effects of different models on the joint extraction task of entity relations in the academic field, and we can draw the following three points:

1) The convolutional neural network layer introduced in series has a better effect on the entity relationship extraction task. In the three sets of comparative experiments on the introduction of convolutional neural networks, the parallel introduction of convolutional neural networks did not improve the overall effect of the model, the series introduction had an improvement of 0.2%, 0.14% and 0.07% in the entity extraction task, and the accuracy rate and F1 value in the relationship extraction task were improved by 3.06% and 0.096%, respectively.

**Table 7** Results of different models on entity-relation joint extraction tasks

model	Entity			Relation		
	Precision	Recall	F1	Precision	Recall	F1
1) BiLSTM-CRF	0.6249	0.4451	0.5196	0.6296	0.4580	0.5303
2) W2c-BiLSTM-CRF	0.7004	0.5814	0.6354	0.7330	0.5972	0.6582
3) W2c-Clue-BiGRU-CRF	0.7796	0.5233	0.6262	0.7693	0.5198	0.6204
4) W2c-Clue-BiLSTM- CRF (our model)	0.6895	0.6089	0.6467	0.7173	0.6113	0.6601
5) W2c-Clue-CNN-CRF	0.5932	0.3420	0.4338	0.5631	0.3229	0.4105

2) The BiGRU model has better accuracy in the joint entity and relation extraction task, and BiLSTM has a better effect on recall. In the two comparative experiments, the model using BiGRU for context extraction achieves the best results of 77.96% and 76.69% for entity and relation extraction, which is significantly higher than that of BiLSTM. However, BiLSTM is better in terms of recall, reaching 61.03% and 62.89%, respectively. On the other hand, BiLSTM is still better in terms of comprehensive effect, with F1 values of 64.83% and 67.67%.

3) The effect of using one-way LSTM for information extraction is obviously not as good as that of bidirectional LSTM. At the same time, the results of this group of experiments also confirm that the effect of series CNN is better than that of parallel CNN.

#### 4.5 Entity and Relation Triples

The final output of the model is the labels corresponding to each word in the texts of our corpus. Actually, the result of information extraction requires the relationship between entities in the text, that is, when we extract relations, we obtain a set of a triple form about entities and relations, which contains both entities and their relations, and also indicates the subject entity and the object entity of the relationship. The triple is also the final form needed for downstream tasks, such as building the model of the knowledge graph. If the downstream task wants to use the model results to build a knowledge graph, the entities in the triple group are node options in the knowledge graph, and the relations are the edges of the knowledge graph. Therefore, we need to define rules that match on the model output results of labels to generate the entity-relation triple. Here we explain the entity pairing rules for our label matching output results.

The results of the sequence labeling model (CRF) need to be combined by certain matching rules to obtain complete triples. Zhang, et al. considered the joint extraction task as a sequence labeling task<sup>[34]</sup> and used “BIO” to label entities, and proposed entity relationship extraction rules. However, this type of rule only matches between normal labels, and cannot solve the problem of matching between one-to-many and many-to-many overlapping relationships. In subsequent related studies, overlapping label rules have been continuously introduced and updated to solve this problem<sup>[35]</sup>. Based on the above research, this paper proposes a new rule to match the entities of the model results.

The “OVE” relationship type introduced in this paper includes two labels, “OVE-1” and “OVE-2”, which assign role information to the labels of overlapping relationship entities, so that



overlapping relationship entities also have the status of issuing actions to match each entity. Based on the our “OVE” labels, the triple matching process is as follows:

1) Determine the subject entity and relation based on the “OVE-1” label. If and only if the location information starts with “B”, and the role information is “1”, the first triple is matched with the entity labeled by “OVE-1”, and go to the next step to find the next triple that relates to an object entity for the “OVE-1” entity.

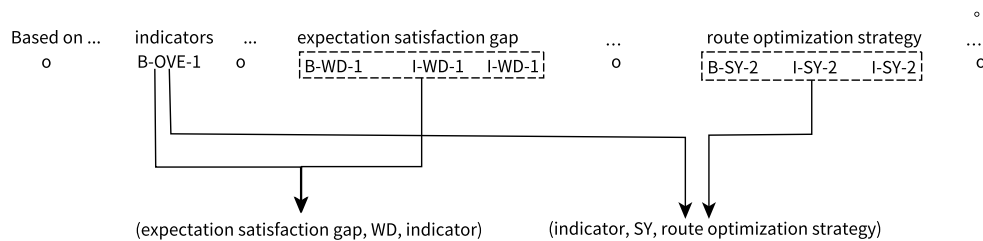
2) Search for an object entity according to the rules to complete the entity matching. Finding an object entity is divided into two cases as follows:

a. If the relationship information contained in the subject entity does not overlap, three pairs are matched in the text: One with the entity of the same relationship, two with the entity labeled “OVE-2” and “OVE-1”. The relationship between two entities is determined by the relationship information contained in the paired primary entity. And each time only the closest entity is paired to form a triple.

b. If the relationship information contained in the object entity overlaps, two pairs are matched in the text: Only pairs with object entities that end with the label “2”, and not with entities that end with the label “OVE-2”. The relationship between the two entities is determined by the relationship information contained in the matching object entity. Only the closest entity is paired to form a triple.

3) Repeat the above steps until all the non-repeating triples are obtained.

Figure 5 shows an example of matching overlapping relations between entities. According to the labels in this sentence we get three complete entities: “indicator”, “expectation satisfaction gap” and “route optimization strategy”. To match all entities and relations, we first find the subject entity and relations, that is, find the entity in which the end of label is “1” and whose relation is annotated. So we get parts of triples: (expectation satisfaction gap, WD, \*) and (indicator, OVE, \*). On the second step, there is an entity labeled with OVE-1, that is, this entity both as a subject entity and object entity, then it is easy to match the two triples obtained in Step 1: Traverse the undetermined triples to pairing based on distance in order. Therefore, filling in (expectation satisfaction gap, WD, \*) we get (expectation satisfaction gap, WD, indicator), and then finding the closest entity labeled with the end of “2” we get “Route optimization strategy”, so the second triple is (indicator, SY, route optimization strategy). After all the primary entities have been matched, we move on to the next sentence.



**Figure 5** Example of matching overlap entities

## 5 Discussion

This paper first constructs a corpus of scientific texts from the academic projects in management science of the National Natural Science Foundation of China, which can reflect the

current academic frontiers in China. Table 8 lists the information of some Chinese corpora built on scientific texts and our dataset. Previous researches are relatively single and simple, and the entity and relation classification is also oversimplistic. For example, Zhao’s data is derived from Chinese Wikipedia, journal articles and dissertations<sup>[27]</sup>, and in their paper they also do not specify the details of the source of articles and dissertations; Zhang’s data is derived from 198 articles from Chinese journal - Journal of the China Society for Scientific and Technical Information<sup>[28]</sup>, Jiang’s corpus uses 25 article texts of Journal of the China Society for Scientific and Technical Information<sup>[5]</sup>. Furthermore, their studies have not carried out the joint extraction of entities and relations. Moreover, none of these previous studies have published their corpus, and we do not have access to their corpora for downstream text analysis work in this field.

**Table 8** Some corpus of Chinese scientific texts

Corpus	Data source	Available
Zhang and Zhang <sup>[28]</sup>	198 articles from Journal of the China Society for Scientific and Technical Information (Chinese)	Not available
Zhao and Wang <sup>[27]</sup>	Chinese wikipedia, journal articles and dissertations (not specified)	Not available
Jiang and Sun <sup>[5]</sup>	25 articles from Journal of the China Society for Scientific and Technical Information (Chinese)	Not available
Our dataset	Texts of projects (2018–2019) from the National Natural Science Foundation of China (management science branch)	Github [ <a href="https://github.com/sobblueayuan/ch_corpus">https://github.com/sobblueayuan/ch_corpus</a> ]

**Table 9** Measurements of some dataset and models

Dataset or model	Language	Entities			Relations		
		<i>P</i>	<i>R</i>	<i>F1</i>	<i>P</i>	<i>R</i>	<i>F1</i>
ScienceIE	English	46.2	48.2	47.1			
SciIE (2020)	English	84.91*	80.94*	82.87*	68.53	65.48	66.97
ACL etc. with MDER 2021	English	76.23	64.2	69.63			
Zhang and Zhang <sup>[28]</sup>	Chinese	69.6	50.46	58.5			
Zhao and Wang <sup>[27]**</sup>	Chinese	83.81	85.24	84.52			
our dataset	Chinese	68.95	60.89	64.67	71.73	61.13	66.01

\*: Extractions of boundaries (Kruiper, etc.); \*\*: Only two tags marked in the dataset.

Table 9 shows the comparison results of entity and relation extraction between our dataset and some other corpus datasets. The English dataset SciIE 2020 performs better in entity extraction<sup>[26]</sup>, while ScienceIE was proposed earlier<sup>[6]</sup> and it is a small dataset that needs to be improved. Hou, et al. proposed a MDER model and improved the effects to achieve 76% in accuracy<sup>[3]</sup>. The effect of Zhao and Wang reached 83.83% in accuracy and 85.52 in F1, and it seems to be at an approximate level of practicality, however, their classification of entity in the dataset included only two tags and is too simple to be used in downstream tasks.

## 5.1 Theoretical Implications

Based on the basic characteristics of scientific texts and the basic elements of literature, and referring to the labeling methods and annotation specifications of English datasets, we give a corresponding annotation specification scheme for Chinese scientific corpus. This normative scheme supports the task of extracting scientific literature information in natural language processing (NLP). The joint extraction of entities and relationships in the academic field strengthens the connection between entities and relationships in the academic field, and improves the limitation of scientific text information extraction research, which is basically only a single type of research.

In the process of building the deep learning model, we discuss the various input features of the model in detail, and stitch the word vector with the manually summarized clue word vector as the initial feature of the model embedding layer. In this paper, the clue word feature is added to the entity relationship extraction model, and the use of this feature makes the accuracy of the model have the highest recall rate and F1 value in a relatively good state, and perform better than the results of other researches in entity extraction of Chinese scientific texts. However, we also have not reached the level of practicality.

In addition, to solve the problem of one-to-many and many-to-one overlapping of entities, we introduce the “OVE” series of labels, and experimental results show that this scheme can improve the identification of overlapping relationship entities.

We have also made improvements to overlapping relationship labels on labels. Based on the research results of predecessors, the overlapping relationship label is improved, and some labels of “OVE-1” and “OVE-2” are added, so that the overlapping relationship entity not only passively waits for pairing to form a triplet, but also can be used as the main role to send actions to form a triplet, which solves the problem of ineffective matching caused by overlapping relationships to some extent.

## 5.2 Practical Implications

This paper builds an academic-oriented professional corpus that can support natural language processing tasks in the academic field. Based on the existing research, we determine the definition of research methods and research objects in the academic field, and propose a label specification scheme for Chinese academic fields, including 6 entity types and 7 relationship types, a total of 29 labels. Then, based on this scheme, the abstracts of the collected scientific literature of the National Natural Science Foundation of China in the field of Chinese management science were manually annotated, and the corpus dataset was created. The dataset currently contains 13,060 entities of different types and 9,728 different relational labeled entities.

This corpus is open for further research by researchers in the field. Although it has not yet reached the application stage, it has a larger sample size than the previous corpus, and the corpus will be open and continuously updated.

## 5.3 Limitations

In general, the performance of these datasets is not yet good enough for large-scale applications and needs to be extended and improved. The corpus of this paper has been published on GitHub and can be used by related work.

The work done in the academic field of this paper is exploratory, and the main work that needs to be improved is as follows:

1) Due to the high cost of manual annotation, the dataset formed in this paper is still small in size, and an effective expansion of the dataset should be carried out within the scope of ability. To improve the efficiency of manual annotation, we have developed a Django-based annotation system, and the next step will be to build an active learning model based on this annotation system to build a labeling tool based on semi-supervised learning, and use this tool to expand our corpus to achieve a corpus based on NSFC project text data from the past 5~10 years.

2) The label classification specification constructed in this paper needs further improvement. The current labels are coarse-grained in terms of relationships, so the entity category information is not fully utilized.

3) The technology in the field of joint extraction of entity relations is constantly updated and advanced, and new cutting-edge technologies can be applied in the follow-up for experiments, and there are many different levels of features for Chinese scientific research literature, such as word segmentation features, syntactic features, positional features, etc., and the features should be fully utilized as much as possible in the model.

## 6 Conclusion

This paper discusses the problem of entity recognition and relationship extraction in natural language processing in the current scientific research field, gives a scheme for entity and relationship classification in information extraction of Chinese scientific texts in management science, and manually annotates the scientific field text corpus collected from the National Science Foundation of China to form a Chinese entity relationship labeling dataset. According to the types and relationships of research objects, tools, techniques or measurements commonly used in academic research, we define six entity types and seven relationship types for entity relationship labeling, use “BIO” labeling schemes for corpus labeling, and obtain a joint extraction corpus of entity relationships in Chinese scientific field as a dataset. Our dataset contains 13,060 entities of different types and 9,728 different relationship labeled entities as a dataset for further research. The joint extraction corpus of entity relationships in the scientific field constructed in this paper has a good effect and can be applied to the downstream tasks of natural language processing in the scientific field.

The deep learning model BiLSTM-CN-CRF is constructed to extract entity relationships in the academic field. In terms of feature input of the model, we introduce the characteristics of clue words, and take the mixed features combined with word vectors as the input of the model, and at the same time, we introduce a CNN layer into the joint model BiLSTM-CRF to strengthen the automatic extraction of local features of the input text. By comparing our model with other types of deep learning models, the experimental results show that in the joint extraction of entity relations in the academic field studied in this paper, the mixed feature input with clue word features and the introduction of convolutional neural network in tandem have a significant effect improvement. In terms of the entity prediction effect, the accuracy rate reached 69.15%, the recall rate reached 61.03%, and the F1 value reached 64.83%. In terms of

relationship prediction effect, the accuracy rate reached 73.24%, the recall rate reached 62.89%, and the F1 value reached 67.67%, which improved the final effect.

The main value of this paper lies in the following three aspects: 1) This paper gives a new entity relationship classification specification for Chinese scientific information extraction based on the English corpus classification and studies in Chinese scientific texts. The introduction of the overlapping relationship series label “OVE” solves the problem of many-to-many relationship overlap to a certain extent. 2) By collecting the 2018–2019 project texts in the management branch of NSFC, this paper builds a corpus dataset for joint extraction of scientific entities and relationships by manual annotation, which can be used for NER follow-up tasks. 3) Based on mixed feature input of word vectors and clue words and BiLSTM-CNN-CRF architecture, a joint extraction model of entity relations is constructed, which can make good predictions and achieve good results.

## References

- [1] Zhang Z, Tam W, Cox A. Towards automated analysis of research methods in library and information science. *Quantitative Science Studies*, 2021, 2(2): 698–732.
- [2] Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 2015, 66(11): 2215–2222.
- [3] Hou L, Zhang J, Wu O, et al. Method and dataset entity mining in scientific literature: A CNN + BiLSTM model with self-attention. *Knowledge-Based Systems*, 2022, 235: 107621.
- [4] Tan Z, Liu C, Mao Y, et al. AceMap: A novel approach towards displaying relationship among academic literatures. *The 25th International Conference Companion on World Wide Web - WWW'16 Companion*, ACM Press, 2016: 437–442.
- [5] Jiang T, Sun J. Learning concept hierarchies from scientific articles for ontology construction. *Journal of the China Society for Scientific and Technical Information*, 2017, 36(10): 1080–1092.
- [6] Luan Y, Ostendorf M, Hajishirzi H. Scientific information extraction with semi-supervised neural tagging. *Association for Computational Linguistics*, 2017: 2641–2651.
- [7] Liu P, Guo Y, Wang F, et al. Chinese named entity recognition: The state of the art. *Neurocomputing*, 2022, 473: 37–53.
- [8] Goyal A, Gupta V, Kumar M. Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, 2018, 29: 21–43.
- [9] Guo X, Zhou H, Su J, et al. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism. *Computers and Electronics in Agriculture*, 2020, 179: 105830.
- [10] Wen Y, Fan C, Chen G, et al. A survey on named entity recognition. *Lecture Notes in Electrical Engineering. Communications, Signal Processing, and Systems*, Springer, 2020, 571: 1803–1810. doi: 10.1007/978-981-13-9409-6\_218.
- [11] Hou Y, Jochim C, Gleize M, et al. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. *The 57th Annual Meeting of the Association for Computational Linguistics*, 2019. <http://arxiv.org/abs/1906.09317>.
- [12] Zhao S, Cai Z, Chen H, et al. Adversarial training based lattice LSTM for Chinese clinical named entity recognition. *Journal of Biomedical Informatics*, 2019, 99: 103290.
- [13] Li Y, Du G, Xiang Y, et al. Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge. *Journal of Biomedical Informatics*, 2020, 106: 103435.
- [14] Liu J, Gao L, Guo S, et al. A hybrid deep-learning approach for complex biochemical named entity recognition. *Knowledge-Based Systems*, 2021, 221: 106958.
- [15] Lerner I, Paris N, Tannier X. Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of Biomedical Informatics*, 2020, 102: 103356.
- [16] Lee C M, Huang C K, Tang K M, et al. Iterative machine-learning chinese term extraction. *The Outreach*

- of Digital Libraries: A Globalized Resource Network. Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, 7634: 309–312. doi: 10.1007/978-3-642-34752-8\_37.
- [17] Geng Q, Deng S, Jia D, et al. Cross-domain ontology construction and alignment from online customer product reviews. *Information Sciences*, 2020, 531: 47–67.
- [18] Nasar Z, Jaffry S W, Malik M K. Information extraction from scientific articles: A survey. *Scientometrics*, 2018, 117(3): 1931–1990.
- [19] Wan Q, Wei L, Chen X, et al. A region-based hypergraph network for joint entity-relation extraction. *Knowledge-Based Systems*, 2021, 228: 107298.
- [20] Geng Z, Zhang Y, Han Y. Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, 2021, 429: 132–140.
- [21] Zhang Y K, Liu M F, Hu H J. Chinese medical entity classification and relationship extraction based on joint neural network model. *Computer Engineering & Science*, 2022. [http://en.cnki.com.cn/Article\\_en/CJFDTotol-JSJK201906021.htm](http://en.cnki.com.cn/Article_en/CJFDTotol-JSJK201906021.htm).
- [22] Liu X, Liu Y, Wu H, et al. A tag based joint extraction model for Chinese medical text. *Computational Biology and Chemistry*, 2021, 93: 107508.
- [23] Zhou Z, Mu W, Yang X, et al. Joint extraction of entities and relations for Chinese text of tea. The 2020 8th International Conference on Information Technology: IoT and Smart City, 2020: 146–152. doi: 10.1145/3446999.3447027.
- [24] Zadeh B Q, Schumann A K. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. The Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016: 1862–1868.
- [25] Yin Z, Wu S, Yin Y, et al. Relation classification in scientific papers based on convolutional neural network. *Lecture Notes in Computer Science*, Springer, 2019: 242–253.
- [26] Kruiper R, Vincent J F V, Chen-Burger J, et al. A scientific information extraction dataset for nature inspired engineering. The 12th Conference on Language Resources and Evaluation, 2020: 2078–2085.
- [27] Zhao H, Wang F. A deep learning model and self-training algorithm for theoretical terms extraction. *Journal of the China Society for Scientific and Technical Information*, 2018, 37(9): 923–938.
- [28] Zhang C, Zhang Y. Automatic recognition of research methods from the full-text of academic articles. *Journal of the China Society for Scientific and Technical Information*, 2020, 39(6): 589–600.
- [29] Zhang C, Xie Y, Song Y. Association analysis of fine-grained knowledge entities in academic texts. *Library Tribune*, 2021, 41(3): 12–20.
- [30] Jiang T, Sun J. Extracting non-taxonomic relationships for ontology learning of scientific resources. *Library and Information Service*, 2016, 60(20): 112–122.
- [31] Zhang C, Mayr P, Lu W, et al. JCDL2022 workshop: Extraction and evaluation of knowledge entities from scientific documents (EEKE2022). The 22nd ACM/IEEE Joint Conference on Digital Libraries, ACM, 2022: 1–2.
- [32] Wang Y, Zhang C, Li K. A review on method entities in the academic literature: Extraction, evaluation, and application. *Scientometrics*, 2022, 127(5): 2479–2520.
- [33] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. *Computer Science*, 2015. <http://arxiv.org/abs/1508.01991>.
- [34] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme. *Association for Computational Linguistics*, 2017: 1227–1236. doi: 10.18653/V1/p17-1113.
- [35] Dai D, Xiao X, Lü Y, et al. Joint extraction of entities and overlapping relations using position-attentive sequence labeling. The AAAI Conference on Artificial Intelligence, 2022: 6300–6308.