Enhancing Ontology Knowledge for Domain-Specific Joint Entity and Relation Extraction

Xiong Xiong[1,2], **Chen Wang**[1,2], **Yunfei Liu**[1,2], **Shengyang Li**[1,2]*
[1]Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{xiongxiong20,wangchen21,liuyunfei,shyli}@csu.ac.cn

Abstract

Pre-trained language models (PLMs) have been widely used in entity and relation extraction methods in recent years. However, due to the semantic gap between general-domain text used for pre-training and domain-specific text, these methods encounter semantic redundancy and domain semantics insufficiency when it comes to domain-specific tasks. To mitigate this issue, we propose a low-cost and effective knowledge-enhanced method to facilitate domain-specific semantics modeling in joint entity and relation extraction. Precisely, we use ontology and entity type descriptions as domain knowledge sources, which are encoded and incorporated into the downstream entity and relation extraction model to improve its understanding of domain-specific information. We construct a dataset called SSUIE-RE for Chinese entity and relation extraction in space science and utilization domain of China Manned Space Engineering, which contains a wealth of domain-specific knowledge. The experimental results on SSUIE-RE demonstrate the effectiveness of our method, achieving a 1.4% absolute improvement in relation F1 score over previous best approach.

1 Introduction

Extracting relational triples from plain text is a fundamental task in information extraction and it's an essential step in knowledge graph (KG) construction (Lin et al., 2015). Traditional methods perform Named Entity Recognition (NER) and Relation Extraction (RE) in a pipelined manner, that is, first extract entities, and then perform relation classification on entity pairs (Zhou et al., 2005; Chan and Roth, 2011; Gormley et al., 2015). However, since the entity model and relation model are modeled separately, pipelined methods suffer from the problem of error propagation. To address this issue, some joint methods have been proposed (Yu and Lam, 2010; Li and Ji, 2014; Zheng et al., 2017; Wang and Lu, 2020; Yan et al., 2021; Xiong et al., 2022). The task of joint entity and relation extraction aims to simultaneously conduct entity recognition and relation classification in an end-to-end manner.

In recent years, with the development of pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018), many entity and relation extraction methods have adopted the paradigm of pre-training and fine-tuning. They utilize PLMs to encode the contextual representations of input text and design various downstream models for task-specific fine-tuning. However, when employed for domain-specific entity and relation extraction, this paradigm suffers from problems of semantic redundancy and insufficiency of domain-specific semantics, particularly in highly specialized domains. On the one hand, PLMs are usually trained on general-domain corpora, which results in a significant amount of redundant semantic information that may not be relevant to specific domains and a lack of sufficient domain-specific semantic information. On the other hand, modeling domain-specific information in this paradigm depends primarily on the role of downstream model and domain-specific labeled data in the fine-tuning stage. However, due to the significantly smaller parameter size of downstream model compared to PLMs and the limited availability of domain-specific labeled data, the effectiveness of domain-specific semantic information modeling is constrained.

*Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Consequently, some methods attempt to incorporate domain knowledge into entity and relation extraction models to enhance the their comprehension of domain-specific information . These methods can be broadly categorized into two groups according to how knowledge is introduced: pre-training domain-specific language models and integrating domain-specific knowledge graph information into models. Methods of domain-specific pre-training utilize large-scale domain corpora to facilitate continuous pre-training of existing general-domain language models (Araci, 2019; Peng et al., 2019; Lee et al., 2020) or, alternatively, to perform domain-specific pre-training from scratch (Chalkidis et al., 2020; Gu et al., 2021). However, in certain specialized domains, there may be a dearth of enough domain-specific corpora to support domain-specific pre-training. Another category of methods involve integrating domain-specific knowledge graph information into models, where entity mentions in input text are linked to the corresponding entities in knowledge graph, and then the relevant information of the linked entities in the knowledge graph is incorporated into models (Lai et al., 2021; Roy and Pan, 2021; Yang et al., 2021; Zhang et al., 2022). Some of these knowledge graph integration methods are designed simply for the task of relation extraction (RE) where the entities in the sentence are pre-specified, rather than the task of joint entity and relation extraction. In addition, a prerequisite for this kind of approaches is the availability of a well-constructed domain-specific knowledge graph, which is scarce and expensive for some highly specialized domains.

In this study, we explore how to incorporate domain knowledge for the task of joint entity and relation extraction in space science and utilization domain of China Manned Space Engineering. Due to the lack of sufficient domain-specific corpora to support the pre-training of large-scale language models and the absence of well-constructed domain-specific knowledge graphs, the aforementioned approaches cannot be directly used for domain knowledge enhancement. We propose an ontology-enhanced joint entity and relation extraction method (**OntoRE**) for space science and utilization domain. The predefined domain-specific ontology involves many highly specialized entity types that interconnected by different semantic relations, which frames the knowledge scope and defines the knowledge structure in this domain, so it is an appropriate source of domain knowledge. The ontology can be formalized as a graph structure containing nodes and edges, where nodes represent entity types and edges represent relation types. Furthermore, drawing inspiration from the manner in which humans comprehend specialized terminology, we add descriptions for each entity type in the ontology to enhance the semantic information of entity types. We serialize the ontology graph and then adopt an ontology encoder to learn the embeddings of ontology knowledge. The encoded ontology features are fused with input sentence features, and then the entity and relation extraction is carried out under the guidance of ontology knowledge. To evaluate our model, we construct a dataset called **SSUIE** (**S**pace **S**cience and **U**tilization **I**nformation **E**xtraction), which contains rich knowledge about space science and utilization in the aerospace field. This work exclusively pertains to the problem of entity and relation extraction, therefore our model was evaluated on the subset of SSUIE specifically designed for entity and relation extraction, namely **SSUIE-RE**.

The main contributions of this work are summarized below:

1. A dataset named SSUIE-RE is proposed for Chinese entity and relation extraction in space science and utilization domain of China Manned Space Engineering. The dataset is enriched with domain-specific knowledge, which contains 19 entity types and 36 relation types.
2. An ontology-enhanced method for domain-specific joint entity and relation extraction is proposed, which substantially enhances domain knowledge without the need of domain knowledge graphs or large-scale domain corpora. Experimental results show that our model outperforms previous state-of-the-art works in terms of relation F1 score.
3. The effect of domain ontology knowledge enhancement is carefully examined. Our supplementary experiments show that the ontology knowledge can improve the extraction of relations with varying degrees of domain specificity. Notably, the benefit of ontology knowledge augmentation is more evident for relations with higher domain specificity.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

2 Related Work

Among the representative entity and relation extraction approaches in recent years, some focus on solving the problem of triple overlapping (Zeng et al., 2018; Nayak and Ng, 2020; Yu et al., 2020; Wei et al., 2020; Wang et al., 2020) and some focus on the problem of task interaction between NER and RE (Wang et al., 2018; Yan et al., 2021; Xiong et al., 2022). However, the challenge of effectively integrating domain knowledge into entity and relation extraction models to improve their applicability in specific fields, has not been solved well by previous works. We survey the representative works on topics that are most relevant to this research: *domain-specific pre-training* and *integrating knowledge graph information*.

Domain-specific pre-training In order to enhance the domain-specific semantics in PLMs, this family of approaches uses domain corpora to either continue the pre-training of existing generic PLMs or pre-train domain-specific language models from scratch. FinBERT (Araci, 2019) is initialized with the standard BERT model (Devlin et al., 2019) and then further pre-trained using financial text. BioBERT (Lee et al., 2020) and BlueBERT (Peng et al., 2019) are further pre-trained from BERT model using biomedical text. Alsentzer et al. (2019) conduct continual pre-training on the basis of BioBERT, and PubMedBERT (Gu et al., 2021) is trained from scratch using purely biomedical text. Chalkidis et al. (2020) have explored both strategies of further pre-training and pre-training from scratch and release a family of BERT models for the legal domain.

Integrating knowledge graph information This category of methods infuse knowledge into the entity and relation extraction models with the help of external knowledge graph. Lai et al. (2021) adopt the biomedical knowledge base *Unified Medical Language System (UMLS)* (Bodenreider, 2004) as the source of knowledge. For each entity, they extract its semantic type, sentence description and relational information from UMLS with an entity mapping tool MetaMap (Aronson and Lang, 2010), and then integrate these information for joint entity and relation extraction from biomedical text. Roy and Pan (2021) fuse UMLS knowledge into BERT model for clinical relation extraction and explore the effect of different fusion stage, knowledge type, knowledge form and knowledge fusion methods. Zhang et al. (2022) integrate the knowledge from Wikidata [0] into a generative framework for relational fact extraction.

To the best of our knowledge, only a limited number of specialized domains can meet the conditions for applying the two aforementioned methods of enhancing domain knowledge, mainly including biomedical, financial, and legal fields. These fields are comparatively prevalent in human life, so there are more likely to be a considerable amount of domain corpora and data in these fields. However, in highly specialized fields like aerospace, both the large-scale domain-specific corpora and well-constructed domain-specific knowledge graph are scarce. Our proposed method only utilize the ontology and entity type descriptions to inject domain knowledge into entity and relation model without additional prerequisites.

3 Method

In this section, we introduce the architecture of OntoRE. As shown in Figure 1, the model mainly includes four parts: knowledge source, knowledge serialization, knowledge encoding and knowledge fusion. In the following subsections, we provide a detailed description of each component.

3.1 Knowledge Source

In the process of human learning professional knowledge, a common practice is to first understand the meaning of each specialized term and then establish the interrelationships between them. Following this pattern, we leverage ontology and entity type descriptions as domain knowledge sources to augment the capacity of entity and relation extraction models to comprehend domain-specific information. Ontology defines the semantic associations among specialized entity types in the domain, while entity type descriptions provide further explanations for each type of specialized terms. For the space science and utilization domain, the ontology is predefined in SSUIE-RE dataset (see Section 4.1). We collect the official descriptions of domain-specific entity types from the *China Manned Space* official website [1].

[0] https://www.wikidata.org
[1] http://www.cmse.gov.cn

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

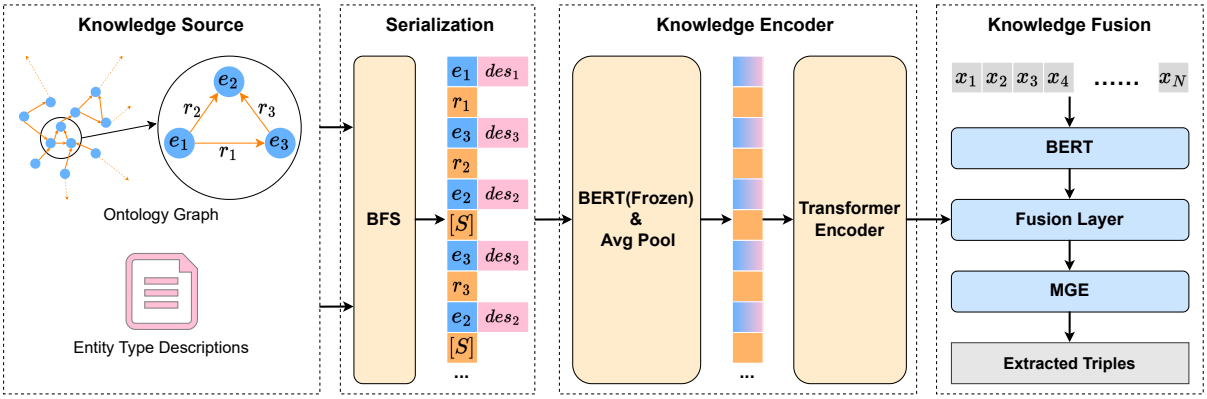

Figure 1: The architecture of the proposed OntoRE framework. We formalize the ontology as a directed graph, where the nodes (blue) represent the predefined entity types and the edges (orange) represent the predefined relation types. The ontology graph is serialized through Breadth First Search (BFS) algorithm. The special marker "$[S]$" represents the end of each level of BFS. des_i denotes the descriptions of entity type e_i. MGE (Xiong et al., 2022) is used as a baseline to verify the effect of our knowledge enhancement method.

Compared to large-scale pre-training corpora and domain-specific knowledge graphs, ontology and entity type descriptions are more accessible for highly specialized domains like space science and utilization.

3.2 Knowledge Serialization

The ontology is a graph structure, where nodes represent entity types and edges represent relation types. It can be formalized as a directed graph $G = (V, E)$, where $V = \{e_1, e_2, \ldots, e_M\}$ denotes the set of predefined entity types and M is the number of predefined entity types. E denotes a multiset of predefined relation types. Additionally, to enrich the semantic information of the entity type nodes in the graph, we append the corresponding entity type descriptions to each node:

$$V' = \{(e_1, des_1), (e_2, des_2), \ldots, (e_M, des_M)\}, \tag{1}$$

where des_m denotes the descriptions of entity type e_m. Then the resulting new graph with the added entity type descriptions can be represented as $G' = (V', E)$.

To facilitate the integration of ontology graph knowledge into entity and relation extraction models that are typically based on sequences, we serialize it using the Breadth First Search (BFS) algorithm while maintaining the structural and semantic properties of the original graph. The graph is represented as an adjacency list in BFS. Before performing the BFS traversal, we initially sort the nodes based on the frequency of their occurrence as entity types in the dataset. Subsequently, we sort the neighboring nodes and edges based on the sum of the node frequency and the edge frequency. This ensures that the nodes and edges with higher frequency in dataset will be traversed first. Then the sorted adjacency list of G is input into the BFS algorithm. During the BFS traversal, we insert a special marker "$[S]$" at the end of each layer of BFS traversal. Taking the nodes e_1, e_2, and e_3 shown in Figure 1 as an example, the first special marker denotes the end of traversing the triple types with e_1 as the head entity, while the second special marker denotes the end of traversing the triple types with e_3 as the head entity, which conveys the structural information among the nodes in the graph. Formally, the BFS serialization process is summarized in Algorithm 1. Then we get an ontology sequence s^π of nodes and edges in the order visited by BFS with level markers:

$$s^\pi = \{s_1^\pi, s_2^\pi, \ldots, s_L^\pi\}, \tag{2}$$

where L represents the length of the ontology sequence obtained by BFS traversal.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Algorithm 1 BFS traversal with level markers

Input: A sorted adjacency list of ontology graph $G' = (V', E)$
Output: A list s^{π} of nodes and edges in the order visited by BFS, with level markers

$s^{\pi} \leftarrow$ empty list
$q \leftarrow$ empty queue
Enqueue the first node of G' into q
Mark all the nodes as unvisited
while q is not empty **do**
  $size \leftarrow$ size of q
  for $i \leftarrow 1$ to $size$ **do**
    $v \leftarrow$ dequeue a node from q
    if v is visited **then**
      break
    end if
    Append v to s^{π}
    for each unvisited neighbor w of v **do**
      Enqueue w into q
      Append the edge (v, w) to s^{π}
      Append w to s^{π}
    end for
    Append a level marker "$[S]$" to s^{π}
    Mark v as visited // The triple types with v as the head entity type have all been traversed
  end for
end while
return s^{π}

3.3 Knowledge Encoding

The elements in s^{π} consist of texts with varying lengths, which encompass relation type words, special markers, and texts formed by concatenating entity type words with their corresponding entity type description words. To get a preliminary semantic representations of these texts, we initialize the representation of each element in s^{π} with a frozen BERT encoder (Devlin et al., 2019) and employ average pooling to unify the feature size. Then we can generate a representation h_k for each element in s^{π} as follows:

$$h_k = \mathrm{AvgPool}(\mathrm{BERT}_{\mathrm{frozen}}(s_k^{\pi})), k \in \{1, 2, \ldots, L\}, \tag{3}$$

where $h_k \in \mathbb{R}^d$ and d is the hidden size of BERT. $\mathrm{AvgPool}(\cdot)$ denotes the operation of average pooling. The representations of the whole ontology sequence s^{π} is concatenated by h_k:

$$H_{s^{\pi}} = [h_1, h_2, \ldots, h_L], \tag{4}$$

where $H_{s^{\pi}} \in \mathbb{R}^{L \times d}$. The feature information in $H_{s^{\pi}}$ are individually encoded from each element in s^{π}. To further capture the inherent information in the ontology sequence, we use a Transformer Encoder (Vaswani et al., 2017) to obtain the final ontology knowledge representations $H_{know} \in \mathbb{R}^{L \times d}$:

$$H_{know} = \mathrm{TransformerEncoder}(H_{s^{\pi}}), \tag{5}$$

3.4 Knowledge Fusion

Given the encoded ontology knowledge representations H_{know}, it can be integrated into different downstream entity and relation extraction models for knowledge enhancement. We select the state-of-the-art methods that have performed best on publicly available benchmark datasets in recent years, and then we evaluate these algorithms on the SSUIE-RE dataset (evaluation results are shown in Table 1). We select the MGE model (Xiong et al., 2022), which performs the best on SSUIE-RE, as our baseline for

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

comparison, and infuse ontology knowledge into it to verify the effectiveness of ontology knowledge enhancement. MGE model uses BERT to encode the contextual information of input sentences, and designs a multi-gate encoder (MGE) based on gating mechanism to filter out undesired information and retain desired information, then performs decoding with table-filling module (Miwa and Sasaki, 2014). We infuse ontology knowledge at the position between the BERT layer and MGE layer, as shown in Figure 1.

We have explored different fusion methods to integrate ontology knowledge representations with input sentence representations, including appending, concatenation and addition. Regarding the appending operation, we concatenate the ontology knowledge representations H_{know} with the input sentence representations along the sequence length dimension. We then apply a self-attention layer to model the guiding effect of ontology knowledge on the extraction of entities and relations from the sentence. The fused representations are calculated as follows:

$$H_{append} = \mathrm{SA}\left([H_b; H_{know}]\right), \tag{6}$$

where $\mathrm{SA}(\cdot)$ means the self-attention layer and H_b denotes the input sentence representations extracted by BERT. $[;]$ denotes the operation of appending, that is, concatenating along the sequence length dimension.

In the case of the concatenation and addition fusion methods, a linear transformation is initially employed to unify the feature dimensions. After this step, the input representations H_b and ontology knowledge representations H_{know} are concatenated along the hidden size dimension or added. The concatenation fusion method can be formalized as below:

$$H_{concat} = \mathrm{Concat}\left(H_b, \mathrm{Linear}(H_{know})\right), \tag{7}$$

where $\mathrm{Concat}(\cdot)$ means the operation of concatenation along the hidden size dimension and $\mathrm{Linear}(\cdot)$ denotes linear transformation. And the fusion method of addition can be formalized as below:

$$H_{add} = H_b + \mathrm{Linear}(H_{know}). \tag{8}$$

Then the representations fused with ontology knowledge is input into the downstream MGE model to obtain the final results of entity and relation extraction.

4 Experiments

4.1 SSUIE-RE Dataset

To evaluate our method, we construct a SSUIE-RE dataset for entity and relation extraction in the space science and utilization domain , which contains rich domain expertise about space science and utilization in the aerospace field. The process of creating SSUIE-RE can be divided into two steps:

Corpora collection and preprocessing The corpora is collected from published professional technical documents in the field, official websites related to China Manned Space Engineering, and Web pages returned by the Google search engine for in-domain professional terms. Before annotation, we preprocess the collected corpora using the following measures:

- We only select Chinese texts and discard texts that are in other languages.
- The invisible characters, spaces and tabs are removed, which are generally meaningless in Chinese.
- In order to eliminate excessively short sentences and incomplete sentences, we split the texts at the Chinese sentence-ending punctuation symbols (e.g., period, question mark, exclamation point), and only retain sentences with more than 10 characters.
- We deduplicate the segmented sentences.

Human annotation We invite annotators with related majors in aerospace field to annotate the processed corpora on the brat[2] platform. The brat platform is an online environment for collaborative text annotation. To ensure the annotation quality, pre-labeling is carried out prior to the formal labeling stage, which aims

[2] https://brat.nlplab.org/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

to ensure that all annotators reach a unified and accurate understanding of the labeling rules. During the annotation process, each sentence is annotated by at least two annotators. If there are inconsistent annotations, the annotation team will discuss the corresponding issue and reach a consensus.

Our final constructed dataset contains 19 entity types, 36 relation types, and 66 triple types. The dataset contains 6926 sentences, 58,771 labeled entities and 30,338 labeled relations. We randomly split the dataset into training (80%), development (10%) and test (10%) set.

4.2 Evaluation and Implementation Details

Following standard evaluation protocol, we use precision (Prec.), recall (Rec.), and micro F1 score (F1) to evaluate our model. The results of NER are considered as correct if the entity boundaries and entity types are both predicted correctly. The results of RE are considered correct if the relation types, entity boundaries and entity types are all predicted correctly.

We use the official implementation of the comparison models to evaluate them on the SSUIE-RE dataset. For fair comparison, we adopt *chinese-bert-wwm* (Cui et al., 2021) as the pre-trained language model for all the models. Our proposed OntoRE model is trained with Adam optimizer for 100 epochs, and the batch size and learning rate are set to be 4 and 2e-5 respectively. The max length of input sentence is set to 300 characters. All the models are trained with a single NVIDIA Titan RTX GPU. The models that achieves the best performance on the development set is selected, and its F1 score on the test set is reported.

4.3 Comparison Models

To ensure a rigorous evaluation, we carefully select state-of-the-art algorithms that have demonstrated superior performance on publicly available benchmark datasets in recent years, and then evaluate their performance on the SSUIE-RE dataset. We compare our model with the following models: (1) **TPLinker** (Wang et al., 2020): this method formulates the task of joint entity and relation extraction as a token pair linking problem, and introduces a handshaking tagging scheme that aligns the boundary tokens of entity pairs for each relation type. (2) **CasRel** (Wei et al., 2020): it models the relations as functions that map subjects to objects rather than discrete labels of entity pairs, allowing for the simultaneous extraction of multiple triples from sentences without the issue of overlapping. (3) **PFN** (Yan et al., 2021): this work utilizes a partition filter encoder to produce task-specific features, which enable effective modeling of inter-task interactions and improve the joint entity and relation extraction performance. (4) **PURE** (Zhong and Chen, 2021): this study constructs two distinct encoders for NER and RE, respectively, and conducts entity and relation extraction in a pipelined manner. (5) **PL-Marker** (Ye et al., 2022): this work is similar to PURE except that it adopts a neighborhood-oriented packing strategy to better model the entity boundary information and a subject-oriented packing strategy to model the interrelation between the same-subject entity pairs. (6) **MGE** (Xiong et al., 2022): this work designs interaction gates to build bidirectional task interaction and task gates to ensure the specificity of task features, based on gating mechanism.

4.4 Main Result

Table 1 shows the comparison of our model OntoRE with other comparison models on SSUIE-RE dataset. As is shown, OntoRE achieves the best results in terms of relation F1 scores. Although PURE achieves the best performance on NER, its relation F1 score is relatively low due to the pipelined architecture which may encounter error accumulation issues. Similarly, PLMarker, which is also a pipelined method, achieves mediocre results on the SSUIE-RE dataset. Among other compared joint methods, MGE achieves the best relation extraction F1 score, and is therefore selected as the baseline model for ontology knowledge injection. As we can see in the table, OntoRE achieves an absolute entity F1 improvement of +0.6% and absolute relation F1 improvement of +1.4% compared to MGE, which indicates that the ontology knowledge injection can enhance the performance of entity and relation extraction in highly specialized domain. Further observation reveals that the models with the best performance on general domain datasets may not perform well in specific professional domains, which reflects the necessity of introducing domain knowledge for entity and relation extraction in specialized fields.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Model	NER			RE		
	Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)
TPLinker (Wang et al., 2020)	77.0	56.0	64.8	65.3	40.8	50.2
CasRel (Wei et al., 2020)	-	-	-	57.8	53.5	55.6
PFN (Yan et al., 2021)	74.9	75.8	75.4	57.8	62.0	59.8
PURE (Zhong and Chen, 2021)	80.5	80.6	**80.6**	55.0	67.4	60.6
PL-Marker (Ye et al., 2022)	80.2	62.6	70.3	55.5	33.4	41.7
MGE (Xiong et al., 2022)	75.8	76.3	76.0	60.0	64.2	62.0
OntoRE (Ours)	75.0	78.3	76.6	62.4	64.5	**63.4**

Table 1: Overall results of different methods on SSUIE-RE Dataset. The results of all comparison models are implemented using official code. We use the same *chinese-bert-wwm* (Cui et al., 2021) pre-trained encoder for all these models. Results of PURE and PL-Marker are reported in single-sentence setting for fair comparison. Results of OntoRE are reported under the utilization of addition fusion method.

4.5 Effect of Domain Knowledge Enhancement

Although our proposed OntoRE achieves the best results on the SSUIE-RE dataset in terms of the overall relation F1 score, in this section, we take a deeper look and further investigate whether OntoRE's integration of domain knowledge essentially improves the model's ability to comprehend domain-specific information.

We observe that the SSUIE-RE dataset includes entity types with varying levels of specialization, ranging from highly specialized entity types (such as *Space Mission*, *Experimental Platform* and *Space Science Field*, etc.) to more general entity types (such as *Person*, *Location*, and *Organisation*, etc.). We refer to the former as in-domain entity types and the latter as general entity types. According to the degree of domain specificity, we categorize 15 out of the 19 entity types defined in the SSUIE-RE dataset as in-domain entity types, and the remaining 4 as general entity types, as shown in Table 2. Based on this categorization, in-domain entities account for 68% of the total entities in the SSUIE-RE dataset, while general entities account for 32%.

Domain Specificity	Entity Types
In-domain (68%)	*Space Mission, Space Station Segment, Space Science Field, Prize, Experimental Platform, Experimental Platform Support System, Experimental System, Experimental System Module, Patent, Criterion, Experimental Project, Experimental Data, Academic Paper, Technical Report, Research Team*
General (32%)	*Organisation, Person, Time, Location*

Table 2: We divide entity types into in-domain and general entity types according to the degree of domain specificity.

To more accurately evaluate OntoRE's ability to understand domain-specific information, we further differentiate the domain specificity of relation types. A triple type defined in the dataset is composed of a head entity type, a relation type, and a tail entity type in the form of (s, r, o). We assess the degree of domain specificity of a relation type by determining whether the head and tail entities it connects are classified as in-domain entity types, as listed in Table 2. Specifically, we consider a relation to be highly domain-specific when both the head and tail entity types are in-domain. If only one of the two entity types is in-domain and the other is general, the corresponding relation is considered to exhibit weaker domain specificity. Furthermore, relations with head and tail entity types are both general rather than in-domain

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

entity types, are considered to exhibit the weakest domain specificity.

We compare our model with the baseline MGE on the performance of recognizing in-domain and general entities, respectively. And for relation extraction, we compare the performance of our model and baseline in extracting relation types with varying degrees of domain specificity. The experimental results are shown in Table 3 and Table 4.

As shown in Table 3, OntoRE outperforms the baseline in recognizing in-domain and general entity types, with a respective improvement of +0.4% and +0.8% in terms of entity F1 score. Table 4 demonstrates that OntoRE obtains an absolute relation F1 score improvement of +0.5%, +1.5% and 1.7% respectively, as the domain specificity of the relation types increases. The experimental results show that OntoRE improves the performance of extracting relation types with varying degrees of domain specificity, and the benefit of ontology knowledge augmentation is more evident for relations with higher domain specificity. This indicates that the incorporation of ontology knowledge appears to be an effective approach for enhancing the model's ability to understand domain-specialized information, while not weakening its understanding of general information.

Entity Type	**Model**	**NER**		
		Prec.(%)	Rec.(%)	F1(%)
In-domain (68%)	Baseline	73.4	73.0	73.2
	OntoRE	72.0	75.3	**73.6 (+0.4)**
General (32%)	Baseline	79.6	81.6	80.6
	OntoRE	79.8	83.0	**81.4 (+0.8)**

Table 3: NER results of in-domain and general entity types on SSUIE-RE test set. In-domain entities account for 68% in the dataset, and general entities account for 32%.

Relation Type	**Model**	**RE**		
		Prec.(%)	Rec.(%)	F1(%)
IDE = 0 (26%)	Baseline	68.5	72.2	70.3
	OntoRE	69.7	71.8	**70.8 (+0.5)**
IDE = 1 (11%)	Baseline	55.5	42.7	48.2
	OntoRE	60.5	42.2	**49.7 (+1.5)**
IDE = 2 (63%)	Baseline	56.8	64.6	60.4
	OntoRE	59.3	65.3	**62.1 (+1.7)**

Table 4: RE results of relation types with varying degrees of domain specificity on SSUIE-RE test set. IDE (In-Domain Entity) represents the number of in-domain entity types contained in a triple type according to ontology definition. The proportions of relations with IDE=0, IDE=1, and IDE=2 in the SSUIE-RE dataset are 26%, 11%, and 63%, respectively.

4.6 Ablation Study

In this section, we conduct ablation study on the SSUIE-RE dataset to examine the effectiveness of our model, specifically with regard to three factors: knowledge source, knowledge fusion method, and the number of knowledge encoder layers.

4.6.1 Knowledge Source and Fusion Method

We put the two factors of knowledge source and knowledge fusion method together for experimental analysis. For the aspect of knowledge source, we remove the entity type descriptions (denoted as *Des* in Table 5) from the complete OntoRE framework to examine the role of entity type descriptions in

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

knowledge enhancement. For knowledge fusion method, we examine the effects of three fusion methods: appending, concatenation and addition.

Table 5 presents a comparison of the experimental results for different combinations of these factors on the SSUIE-RE dataset. The experimental results show that, under the fusion methods of appending and concatenation, the incorporation of entity type descriptions improves NER F1 scores by 1.4% and 1.4% respectively. However, under the addition fusion method, there is a slight decrease in NER F1 score. This can be attributed to the need for compressing the dimension of the entity type description tensor to match the input sentence tensor before addition, leading to information loss. Across all three fusion methods, the inclusion of entity type descriptions consistently improve the relation F1 scores. Additionally, when using the same combination of knowledge sources, the performance of the appending and concatenation fusion methods is comparable, while the addition fusion method achieves the best relation F1 score. This suggests that the optimal approach is to employ ontology and entity type descriptions as knowledge sources and use the addition fusion method to integrate knowledge representations into the model.

Knowledge Source	**Fusion Method**	**NER**			**RE**		
		Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)
Ontology	Append	71.3	77.9	74.5	61.0	62.2	61.6
Ontology	Concat.	72.9	78.3	75.5	61.3	64.4	62.8
Ontology	Add	74.2	79.4	76.7	62.2	64.4	63.3
Ontology + *Des*	Append	73.3	78.8	75.9	60.6	65.4	62.9
Ontology + *Des*	Concat.	75.7	78.0	76.9	63.0	63.0	63.0
Ontology + *Des*	Add	75.0	78.3	76.6	62.4	64.5	**63.4**

Table 5: Ablation study on SSUIE-RE for knowledge source and knowledge fusion method. *Des* denotes entity type descriptions.

4.6.2 Number of Knowledge Encoder Layers

In the knowledge encoding stage, we utilize Transformer encoder to encode the serialized ontology knowledge, as described in Section 3.3. We conduct ablation study to investigate whether stacking multiple layers of encoders could improve the model performance. Considering the parameter size of Transformer encoder, we only experiment with encoder layers up to three. As shown in Table 6, using two layers of Transformer encoders achieved the best performance (which we employed in our final model), and further stacking of encoders does not result in additional performance improvements.

Knowledge Encoder Layers	**NER**			**RE**		
	Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)
L = 1	70.6	80.8	75.3	58.4	64.9	61.5
L = 2	75.0	78.3	76.6	62.4	64.5	**63.4**
L = 3	75.0	77.7	76.4	61.1	62.0	61.5

Table 6: Ablation study on SSUIE-RE for the number of knowledge encoder layers.

5 Conclusion

In this work, we propose an ontology-enhanced method for joint entity and relation extraction in space science and utilization domain. Our model utilizes ontology and entity type descriptions as sources of domain knowledge, and incorporate them into downstream model to enhance model's comprehension of domain-specific information. We introduce a new dataset, SSUIE-RE, which contains rich domain-specialized knowledge. Experimental results on SSUIE-RE demonstrate that our approach outperforms previous state-of-the-art methods. Moreover, we conduct a detailed analysis of the extraction of entities

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

and relations with different degrees of domain specificity and validate the effectiveness of ontology knowledge enhancement. Overall, our proposed method provides a promising direction for improving the performance of entity and relation extraction in specialized domains with limited resources. In the future, we would like to further explore how to generalize the ontology knowledge enhancement idea to other domain-specific information extraction tasks.

Acknowledgements

This work was supported by the National Defense Science and Technology Key Laboratory Fund Project of the Chinese Academy of Sciences: Space Science and Application of Big Data Knowledge Graph Construction and Intelligent Application Research and Manned Space Engineering Project: Research on Technology and Method of Engineering Big Data Knowledge Mining.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November. Association for Computational Linguistics.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 551–560, Portland, Oregon, USA, June. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE Transactions on Audio, Speech and Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784, Lisbon, Portugal, September. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6248–6260, Online, August. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, Baltimore, Maryland, June. Association for Computational Linguistics.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, Doha, Qatar, October. Association for Computational Linguistics.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of AAAI*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online, November. Association for Computational Linguistics.

Shaolei Wang, Yue Zhang, Wanxiang Che, and Ting Liu. 2018. Joint extraction of entities and relations based on a novel graph scheme. In *IJCAI*, pages 4461–4467. Yokohama.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online, July. Association for Computational Linguistics.

Xiong Xiong, Liu Yunfei, Liu Anqi, Gong Shuai, and Li Shengyang. 2022. A multi-gate encoder for joint entity and relation extraction. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 848–860, Nanchang, China, October. Chinese Information Processing Society of China.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 185–197, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 987–991, Online, August. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland, May. Association for Computational Linguistics.

Xiaofeng Yu and Wai Lam. 2010. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In *Coling 2010: Posters*, pages 1399–1407, Beijing, China, August. Coling 2010 Organizing Committee.

Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Yubin Wang, Tingwen Liu, Bin Wang, and Sujian Li. 2020. Joint extraction of entities and relations based on a novel decomposition strategy. In *Proceedings of ECAI*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia, July. Association for Computational Linguistics.

Sheng Zhang, Patrick Ng, Zhiguo Wang, and Bing Xiang. 2022. Reknow: Enhanced knowledge for joint entity and relation extraction. In *NAACL 2022 Workshop on SUKI*.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada, July. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June. Association for Computational Linguistics.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.