Extracting ISA Relations of Concepts from Books via Weakly Supervised Learning

Haonian Wang
School of Artificial Intelligence, Beijing Normal University
Beijing, China
aiwhn@mail.bnu.edu.cn

Xinyu Tang
School of Artificial Intelligence, Beijing Normal University
Beijing, China
txy@mail.bnu.edu.cn

Yurou Liu
School of Artificial Intelligence, Beijing Normal University
Beijing, China
yrliu@mail.bnu.edu.cn

Zhichun Wang
School of Artificial Intelligence, Beijing Normal University
Beijing, China
zcwang@bnu.edu.cn

ABSTRACT

Relation extraction is an important task in natural language processing. In recent years, researchers have gradually discovered that relation extraction can be used for teaching and learning assessment. Among various dependencies, the ISA is a relation that accounts for a relatively large proportion and is of great significance for many related tasks. However, there is currently little research on the ISA relation extraction between cross-chapter concepts for few-shot in textbooks. In this paper, we propose a distant supervised few-shot relation extraction model based on example construction and self-attention optimization. We use the example construction method in contrastive learning to expand the samples and self-attention mechanism to weight the sentences in the sentence bag to extract the relations of the concepts. Experimental results show that the model we constructed performs better than the previous methods.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

Weakly Supervised Learning, Relation Extraction, Few-shot Learning, Self-attention.

1 INTRODUCTION

Relation extraction refers to the extraction of a triplet (subject, relation, object) from a piece of text to automatically identify a certain semantic relation between entities, which can provide effective support for downstream tasks such as automatic construction of knowledge graphs, search engines, and questions and answers. In recent years, researchers have gradually discovered that relation extraction can be used for teaching and learning assessment, but there is little work related to organizing knowledge specifically for educational purposes.

Typically, learning resources in the field of education will explain multiple knowledge concepts, and the knowledge concepts in one field usually need to be learned step by step, following the order from simple to complex, from abstract to concrete[1]. This order depends on the dependencies between concepts. Among various dependencies, the ISA is a relation that accounts for a relatively large proportion and has a greater influence on the judgment of the learning order. For example, for the concept pairs of Neural Network and Convolutional Neural Network, Neural Network is the parent class, Convolutional Neural Network is the subclass, in the actual learning order we need to learn the knowledge concept of Neural Network first and then further learn Convolutional Neural Network. It is very difficult for learners to master a scientific learning order of knowledge in a completely unfamiliar domain. Figure 1 shows an example of ISA relations extracted in the deep learning domain, where each node is a deep learning concept such as "ANN" and "autoencoder", and the links indicate ISA relations related to these concepts (from superclass to subclass).

Recently, large-scale pre-trained language models such as BERT have been successfully applied in relation extraction. However, such models usually need large amount of training samples. Due to the various size of different concepts, there probably exists diverse training degrees, which is likely to cause overfitting. Meanwhile, in previous work, all sentences containing the concept of knowledge were used as samples, but some of them could not represent the concepts, which lacked information on most of the concepts themselves, and the problem about the weight of the sentences (which sentence is more important to represent the ISA concept relation) is overlooked. Therefore, in our study, we manage to explore a method to solve the problem mentioned above and compare our experimental results with the results of the traditional model.

There are many education resources from which we could achieve relation extraction, but we focus on textbooks because they often provide a structured list of concepts and are often used as the primary educational resources in colleges and universities[2]. Structural information such as textbook catalogs and concept lists is very useful for identifying conceptual relations, and we believe that the model we propose can be easily generalized to other educational resources with structured information.

In a nutshell, our contributions include:

- A distant supervised few-shot relation extraction model based on self-attention mechanism, which is better than the traditional BERT relation extraction model.
- A new method for few-shot relation extraction which uses the example construction method in contrastive learning to expand the dataset and uses the self-attention mechanism to weight the sentences in the sentence bag. This method can be generalized to other few-shot relation extraction problems with structured information.

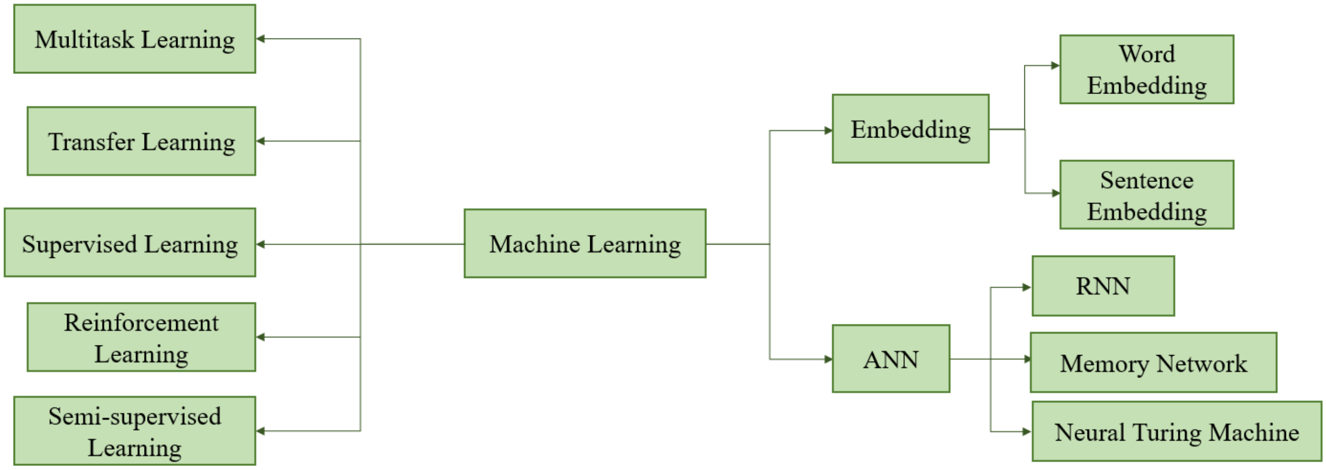


Figure 1: Example of an extracted concept map in deep learning.

The structure of this article is as follows. Section 2 reviews the related work. Section 3 details the distant supervised few-shot relation extraction model based on self-attention mechanism and sample construction method. Section 4 describes the dataset and preparation techniques as well as our experimental results and analysis. Section 5 describes the specific applications. Section 6 is the conclusion and future work.

2 RELATED WORK

The ISA relation between concepts greatly affects the learning order of knowledge and the reading order of teaching resources. Its essence is a kind of relation extraction work.

The area that researchers pay most attention to is to extract the prerequisite relations between concepts. Before 2020, researchers generally used the method of manual feature labeling to extract prerequisite relations. C Liang and J Ye et al. used a rigorous mathematical closure calculation method for relation extraction[3]. S Wang and B Pursel et al. used active learning to design classifiers to filter out the most valuable labels to reduce the amount of training data[4]. W Lu and Y Zhou et al. proposed a domain-specific concept extraction method (iPRL) that enabled the learning of prerequisite relations between concepts without manual labeling of data and gradually improved their performance as the models interact with each other[5]. A common problem with these methods is that they all require experts in the relevant field to manually design the features of the concept pairs, which is not only expensive and inefficient, but also makes the acquisition of features subjective and increases the correlation error.

In order to solve the above problem, researchers have gradually applied machine learning and deep learning to the extraction of conceptual prerequisite relations since 2020. H Yu and H Li et al. used CNN to extract non-superstructured relations in unstructured texts[6]. S Yang and M Zhu et al. used a three-dimensional CNN to convert the students' Q&A content into vectors, put them into LSTM, and updated the portal integration with GRU[7].Y Bai and Y Zhang et al. used the BERT model to extract the prerequisite relations in Wikipedia and presented Chinese datasets annotated with the prerequisite relations[8]. However, we find that these efforts are all for Mooc, Wikipedia or video playlists while textbooks with special structured information and as the main educational resources of universities have been ignored.

In this paper, we propose a weak supervised learning method for ISA relation extraction between concepts in textbooks. We use the example construction method in contrastive learning about sample expanding, and use distant supervised learning to label a very small number of samples. At the same time, we use the self-attention mechanism to weight the sentences in the sentence bag to extract the relations of the concepts. Experimental results show that our method is superior to the previous optimal models.

3 METHODOLOGY

In this paper, we introduce a new method for few-shot relation extraction which use the sample construction method to expand the data and utilize the self-attention mechanism to weight the sentences in the sentence bag. A formal description of our proposed model is explained as follows. The structure of our model is shown in Figure 2. The model consists of three parts: input layer, middle layer and output layer. The input to the model is the concept and its sentence bag. In the middle layer, the sentences in the sentence bag get the vectors through the BERT encoder, and the self-attention mechanism is used to weight the sentences in the sentence bag to get the word vector of each concept. The output layer of the model is whether the two concepts have the ISA relation, which is the classifier to determine whether the two word vectors have.

3.1 Input Embedding

3.1.1 Token Embedding . For each input sequence x, we use BERT Tokenizer to split it into several tokens: $x = (t_1, t_2, \dots, CON \dots, t_L)$,

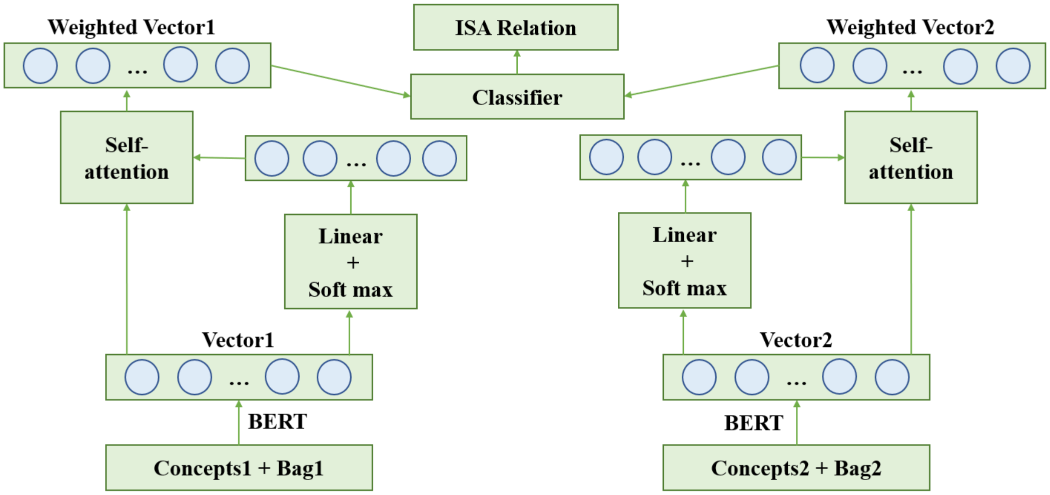


Figure 2: The structure of self-attention Model.

where *CON* is the concept, and L is the length of input sentences. Simultaneously, we mark the beginning and the end of sentences[9] by adding the two special tokens([CLS] and [SEP]). Also, we utilize special token([PAD]) to fill in the blanks at the end of the sentences.

However, concept is more important than other words t_i in relation extraction tasks. So we follow the previous practice which add two special tokens in each sentence ([CON-CLS]) and ([CON-SEP]) to mark the beginning and the end of the concept[10].

3.1.2 Position embedding . Position embedding is used to model the temporal characteristics of positions. It has three characteristics: translation invariance, monotonicity and symmetry.[11]. Therefore, position embedding is put into the input. At the same time, we can add the position embedding and the token embedding because they have the same dimensions.

3.2 Relation Extraction Model

Firstly, we input the embeddings above into the BERT model, and we can get the [CLS] vector of the sentence and hidden state vector of the last layer.

Secondly, we form the one-hot encoding of the concept to increase its weight, multiply it by the hidden state vector and map the result to the sentence dimension to get the mean vector of the concept.

Lastly, following the standard practice, we concatenate the mean vector of concept with [CLS] vector after tanh activate function[12]. And the final result can be obtained through fully connected layer and normalization.

$$\mathbf{logits} = \mathbf{W}(\alpha(\ [\mathbf{c}\ ||\ \tanh(\mathbf{A})]\ )) + \mathbf{B} \tag{1}$$

|| means the concatenation operation, **W**, **B**, α, **c** represent Weight matrix, bias matrix, Nonlinear function and [CLS] vector respectively.

3.3 Distant Supervision

For the task of relation extraction, manual annotation of data may lead to inconsistent annotation results, excessive human resources consumption and the accuracy problem. Therefore, we introduce a distant supervised learning annotation method based on wikidata. Wikidata, which uses a simple data model to store structured data other than text labels and language links, is a collaborative editable knowledge base that provides accurate datasets for our annotation tasks.

In Wikipedia, we specifically consider 9 relations: ISA, part of, instance of, different of, has part or parts, common category, has use, follows, facet of. However, according to the relation of specific concepts in the book, we find that the relation extraction task of multiple relations has the following problems:

- Many concepts have multiple relations, but the importance of different relations is difficult to judge.
- The relation between some concepts is obscure, even manual annotation is difficult to identify.
- Among all the relations of concepts, the relation of ISA accounts for 54% of all relations, while others account for a relatively small proportion, which makes it difficult for the model to identify some relations that rarely appear.

Hence, in this article, we only take ISA relations into consideration.

3.4 Sentence Bag

In previous methods, people often select a sentence containing two or more concepts to judge the relation between them. Unfortunately, in the dataset we selected, the number of annotations is limited, which is not suitable for the relation extraction task.

In this article, we use the method of obtaining the description and definition sentence of two concepts to judge the relation between them. It can complete the task of relation extraction by extracting the semantic information of two words. We adopt the following definition extraction method.

Our goal is to create a sentence bag for each concept that contains 8 key sentences to represent it.

- First, We find the sentences that are definition sentences, or sentences that explain or describe the concept. E.g.:A typical example of a **deep learning** model is a feed-forward deep network or a multilayer perceptron. If the number of sentences is greater than 8, we give priority to the first sentences according to the order in which the concept appears because we consider that the more sentences appear in the front of the book, the more explanatory information it contains about the concepts. Conversely, if the number of sentences is less than 8, we extract sentences with concepts as subjects.
- After completing the above operations, if the number of sentences is still less than 8, we use the method of sample construction(which will be mentioned in 3.5) to enhance the data.

3.5 Sample Construction

We introduce three sample construction methods for data augmentation, namely random delete, random substitute, concept substitute. The detailed methods are as follows.

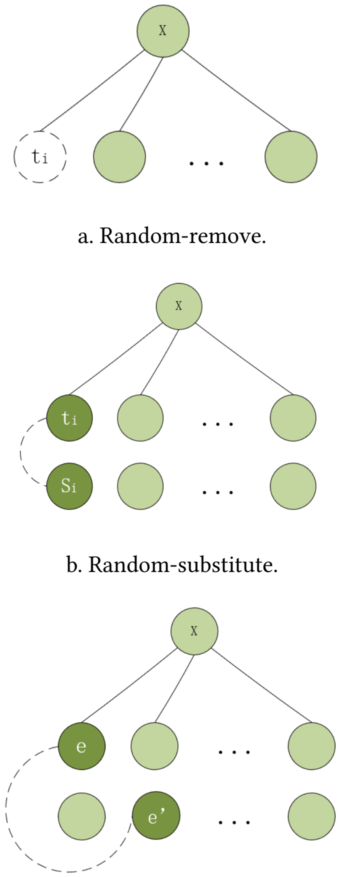
3.5.1 Random remove. A reliable example construction scheme is to select a sentence $x = \{t_1, t_2, \ldots, e, \ldots, t_L\}$ in the sentence bag, which is shown in Figure 3(a). Randomly delete the non-concept t_i in the sentence and get x', ensure $x \neq x'$. It can complete the number of sentence bags in the simplest way with small number of samples, and quickly improve the self-attention method to obtain the weight of the definition sentence (mentioned in section 3.6).

3.5.2 Random substitute. Another reliable way is to randomly replace the words in the sentence, which is shown in Figure 3(b). We randomly replace the non-concept t_i with s_i, get:

$$x' = \{t_1, t_2, \ldots s_i \ldots, e, \ldots, t_L\} \tag{2}$$

This method does not destroy the semantic information in the sentence, and adds many reliable samples.

3.5.3 Concept substitute. The last way is to change the concepts in the sentences, which is shown in Figure 3(c). We consider that the sentences in the model contain rich semantic information. If we only change the concepts, we will not change the semantic information in the whole sentence. For a sentence $x = \{t_1, t_2, \ldots, e, \ldots, t_L\}$ in the sentence bag, We randomly replace e with other concept e' in another sentence bag and keep the semantic information of the whole sentence unchangeable. In this way, the sentence itself is not affected by the specific semantics of concepts, and the semantic information in the sentence can be extracted purely.



a. Random-remove.

b. Random-substitute.

c. Concept-substitute.

Figure 3: Methods for sample construction.

3.6 Self-attention Model

Firstly, we use the definition extraction algorithm (mentioned in Section 3.4) to get a concept sentence bag containing N sentences and use BERT model to encode them respectively.

Secondly, Map the result of the relation extraction model (mentioned in Section 3.2) to a value through a simple fully-connected layer, and use activate function *softmax* to map the value to a conditional probability distribution.

$$\mathbf{z_i} = \mathbf{W_0}\mathbf{s_i} + \mathbf{B_0} \tag{3}$$

$$p_i = softmax(z_i) = \frac{e^{z_i}}{\sum\limits_{j=1}^{N} e^{z_j}} \tag{4}$$

z_i, $\mathbf{W_0}$, $\mathbf{s_i}$, $\mathbf{B_0}$, p_i represent the mapped value, nonlinear weight matrix, sentence vector, bias matrix and probability respectively.

Thirdly, the results are used to calculate the weighted sum of N sentences and we can get the word embedding of the concept through the self-attention mechanism.

$$\mathbf{w_i} = \sum_{k=1}^{N} p_k \times \mathbf{s_k} \tag{5}$$

Here, $\mathbf{w_i}$ represents the word embedding of the concept. p_k represents the weight of the sentence in the sentence bag. $\mathbf{s_k}$ represents sentence vector, to be more specific, it is the $[CLS]$ vector of sentence.

Finally, we concatenate two concept vectors which will enter the fully connected network and use a nonlinear classifier to determine if they have a ISA relation.

$$\mathbf{logits} = \alpha(\mathbf{W_1}[\mathbf{w_i} \mid\mid \mathbf{w_j}] + \mathbf{B_1}) \tag{6}$$

The || represents the concatenation operation, $\mathbf{w_i}$ and $\mathbf{w_j}$ represent two concept vectors, $\mathbf{W_1}$ and $\mathbf{B_1}$ represent weight matrix and bias matrix of the fully connected network, α represents nonlinear classifier.

4 EXPERIMENTS

4.1 Dataset

We use the ***Deep Learning*** book[1] in our experiments. It is a systematic textbook about deep learning and contains totally 885 concepts related to deep learning, which is conducive to extract ISA relations. However, there exists many concepts that have the following problems, which is difficult to extract ISA relation based on this type of concepts:

- The meaning of some concepts is excessively board, such as **Feature** and **Dependency**. The professional relevance with deep learning of these concepts is not strong enough.
- Some concepts appear too few times, whose relevance with other concepts is weak.

For the problem mentioned above, we reference Wikipedia and remove the words which have excessively multiple meanings and low appearing frequency. Finally, we choose 152 concepts which are included by at least 8 key sentences from 885 concepts in this book as our few-shot dataset.

We extract the concepts in the appendix and obtain all the sentences where the concepts are located. In order to reduce the noise in the text, we delete the annotation and chapter identifier in text format, and replace the mathematical formula with the identifier [UNK] which is a special token in BERT. Besides, the original English concept name in the sentence is translated into Chinese. We can gain the key concept sentences with less noise after preprocessing.

4.2 Parameter settings

We use the pretrained BERT-base-Chinese model for the text encoder. After filtering data annotations in wikidata, we extract ISA relation from sentence bag (ISA or not). Finally, we finetune the pretrained BERT-base-Chinese model on the proposed deep-learning dataset with optimizer Adam. We have a total of 40 training epochs. The initial learning rate is set as 1e-5, dropout as 0.3 and weight decay as 1e-8. To make the training more effective, we use the cross-entropy loss.

4.3 Result

For each domain, we apply 5-fold cross validation to evaluate the performance of the proposed method. Our methods will be renamed as:

- RRS(random-remove self-attention(Ours))
- RSS(random-substitute self-attention(Ours))
- CSS (concept substitution self-attention(Ours))

In our experiments, we employ four other methods to compare with our model, including:

- PURE(SOTA's method mentioned in 3.2)[12]
- MP (replace self-attention with Max pooling)
- AP (replace self-attention with average pooling)
- DV (add the difference between two word vectors in the last layer of classifier)

We measure the performance of classifiers on our dataset in terms of the F1-scores. The detailed comparision of different models are as follows.

Table 1: Comparison of the effects of each model.

method	f1-score
RRS	0.740
RSS	0.780
CSS	0.762
PURE	0.734
MP	0.754
AP	0.724
DV	0.724

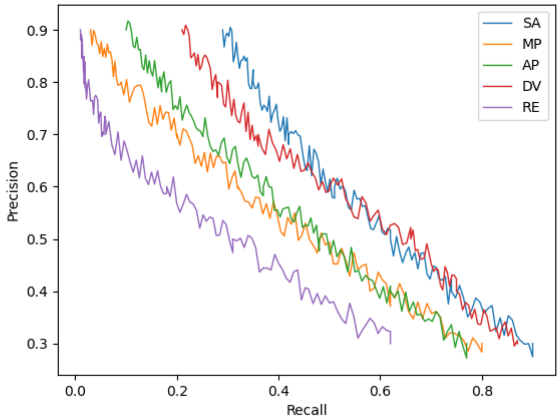


Figure 4: PR-curves of our model and other baselines on deep learning dataset.

We summarize the model performances of our method and baseline models in Table 1. According to the results, we can observe that:

- On the dataset, our proposed methods (RRS, RSS, CSS) achieve the best performance taking the f1-score as the standard.
- RSS method perform better than others in our dataset.

We speculate that, compared with RRS and CSS, the method of synonymous replacement of other components in a sentence can

[1] https://github.com/exacity/deeplearningbook-chinese

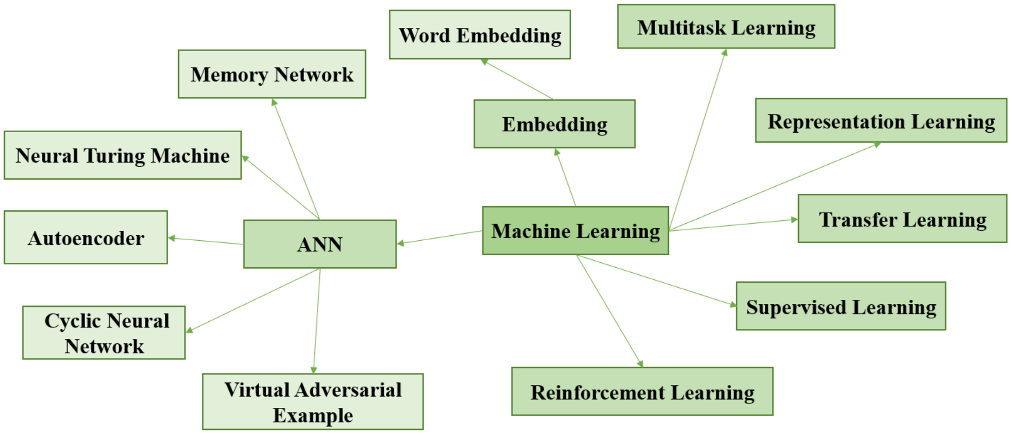


Figure 5: ISA Diagram for Machine Learning.

maintain the original meaning of the sentence and expand the data scale as much as possible, so RSS is a better choice.

Figure 4 [2] shows the precision and recall of the five models. From the curve, we can observe that:

- Compared to PR-curves of other baseline models, our method shifts up the curve a lot.
- Although the curve initially fluctuates violently, our model are basically stabilized during the training process.
- Our methods RRS, RSS and CSS surpass previous PURE(SOTA) model in all ranges along the curve, and they are more balanced between precision and recall.

We infer that in our model, some sentences in the bag describe and define the concept from different dimensions, and the model cannot judge the weight of these sentences well at the beginning. But in the training process, the problem mentioned above tends to be stable and the performance of the model is getting increasingly better. Furthermore, as the SOTA scheme of relation extraction, PURE fails to achieve competitive results. This is because PURE relies on massively labeled data for pre-training, and noisy labels in PURE may influence its model performance.

4.4 Case of study

Our model has the following two applications: ISA relation of concepts and weights of definition sentences.

4.4.1 ISA relation of concepts. In our model, we predict the ISA relations among all concepts in the book and get good results. It is constructive for readers to determine the learning order of this book and plays a good auxiliary effect for reading.

As shown in the Figure 5, the darker the color of the concept block, the higher the level. We take machine learning as an example, artificial neural network, representation learning, etc. are the results of the first layer of divergence, while artificial neural network divergence can include concepts such as cyclic neural network, memory network, etc.

4.4.2 Weights of definition sentences. After weighting by self-attention mechanism, each sentence in the sentence bag will get diverse weights. The more sentences that can express concepts (that is, the sentences we think of as definition sentences and concept description sentences), the greater the weight in the sentence bag. The following is based on the weighting in the sentence bag. This can help readers quickly locate the definition sentence and understand the meaning of the concept.

The results of weighting the sentences of concept Multi-task Learning and Long Short-term Memory were shown in Figure 6 and Figure 7. The ability to describe and characterize concepts from top to bottom decreases accordingly, so the weight obtained in the model decreases from top to bottom.

5 CONCLUSION AND FUTURE WORK

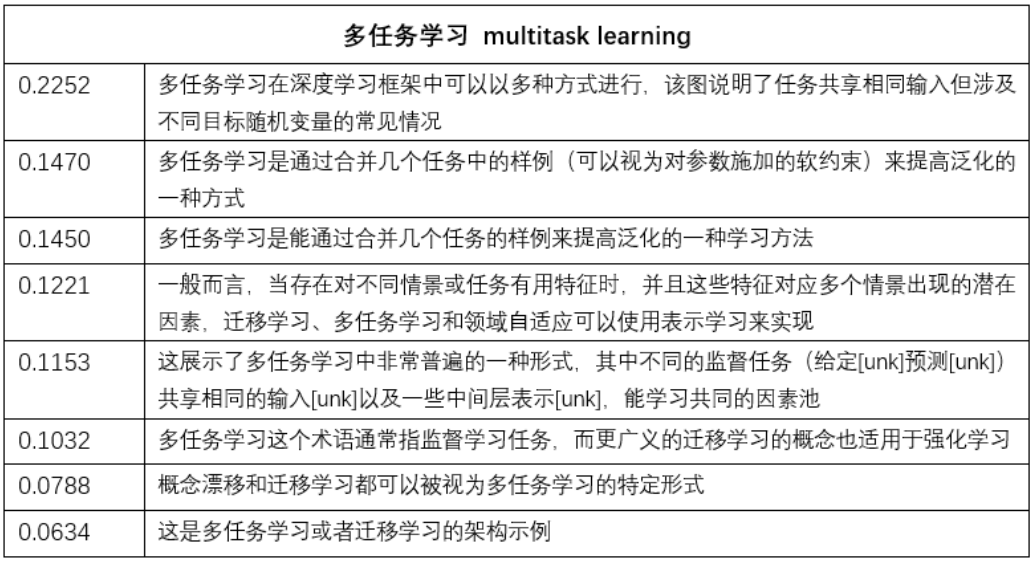
In this paper, we propose a new conceptual ISA relation extraction method, which combines the characteristics of the sentence embedding in the article, the self-attention mechanism feature of the sentence bag and the sample construction method to enhance the data. We use the deep learning dataset to prove the effectiveness of our method. Experimental results show that our method has achieved the best results compared to other traditional methods. In addition, we evaluate the application value of the model.

In the future, we plan to introduce a complete framework of contrastive learning for pre-training of models. In addition, since wikidata contains more types of relations, our next research focuses on how to extract more kinds of relations through the unsupervised learning method of contrastive learning. Simultaneously, we plan to use more datasets in different languages to enhance the robustness of the model.

ACKNOWLEDGEMENT

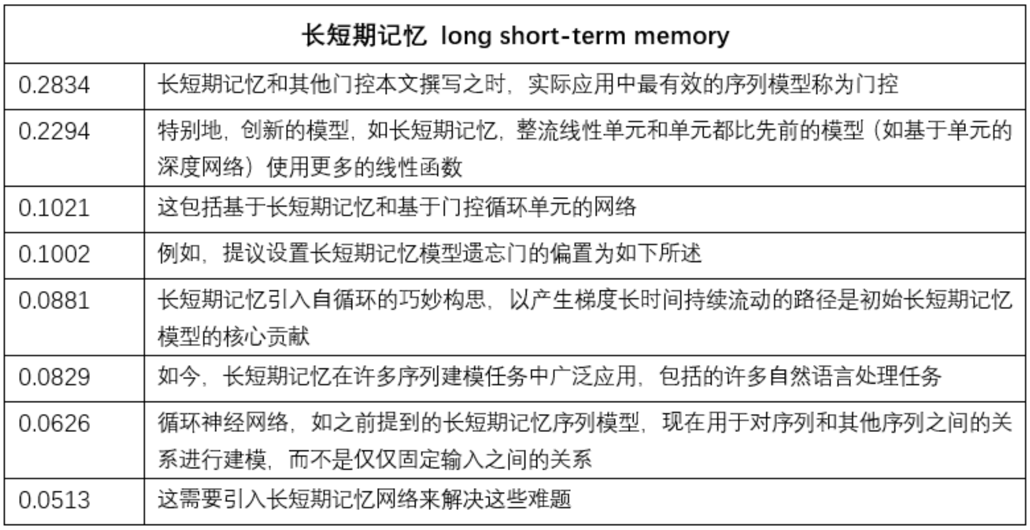
This work is supported by the National Key Research and Development Program of China (No. 2021ZD0113004).

[2] In this figure, SA means our self-attention model, RE means the PURE method



多任务学习 multitask learning	
0.2252	多任务学习在深度学习框架中可以以多种方式进行，该图说明了任务共享相同输入但涉及不同目标随机变量的常见情况
0.1470	多任务学习是通过合并几个任务中的样例（可以视为对参数施加的软约束）来提高泛化的一种方式
0.1450	多任务学习是能通过合并几个任务的样例来提高泛化的一种学习方法
0.1221	一般而言，当存在对不同情景或任务有用特征时，并且这些特征对应多个情景出现的潜在因素，迁移学习、多任务学习和领域自适应可以使用表示学习来实现
0.1153	这展示了多任务学习中非常普遍的一种形式，其中不同的监督任务（给定[unk]预测[unk]）共享相同的输入[unk]以及一些中间层表示[unk]，能学习共同的因素池
0.1032	多任务学习这个术语通常指监督学习任务，而更广义的迁移学习的概念也适用于强化学习
0.0788	概念漂移和迁移学习都可以被视为多任务学习的特定形式
0.0634	这是多任务学习或者迁移学习的架构示例

Figure 6: Sentence bag weights of Multi-task Learning.



长短期记忆 long short-term memory	
0.2834	长短期记忆和其他门控本文撰写之时，实际应用中最有效的序列模型称为门控
0.2294	特别地，创新的模型，如长短期记忆，整流线性单元和单元都比先前的模型（如基于单元的深度网络）使用更多的线性函数
0.1021	这包括基于长短期记忆和基于门控循环单元的网络
0.1002	例如，提议设置长短期记忆模型遗忘门的偏置为如下所述
0.0881	长短期记忆引入自循环的巧妙构思，以产生梯度长时间持续流动的路径是初始长短期记忆模型的核心贡献
0.0829	如今，长短期记忆在许多序列建模任务中广泛应用，包括的许多自然语言处理任务
0.0626	循环神经网络，如之前提到的长短期记忆序列模型，现在用于对序列和其他序列之间的关系进行建模，而不是仅仅固定输入之间的关系
0.0513	这需要引入长短期记忆网络来解决这些难题

Figure 7: Sentence bag weights of Long Short-term Memory.

REFERENCES

[1] Bai Y, Zhang Y, Xiao K, et al. A BERT-Based Approach for Extracting Prerequisite Relations among Wikipedia Concepts[J]. Mathematical Problems in Engineering, 2021.

[2] Wang S, Ororbia A, Wu Z, et al. Using prerequisites to extract concept maps fromtextbooks[C]//Proceedings of the 25th acm international on conference on information and knowledge management. 2016: 317-326.

[3] Liang C, Ye J, Zhao H, et al. Active learning of strict partial orders: A case study on concept prerequisite relations[J]. arXiv preprint arXiv:1801.06481, 2018.

[4] Liang C, Ye J, Wang S, et al. Investigating active learning for concept prerequisite learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[5] Lu W, Zhou Y, Yu J, et al. Concept extraction and prerequisite relation learning from educational data[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 9678-9685.

[6] Zhou Y, Xiao K. Extracting prerequisite relations among concepts in wikipedia[C]//2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019: 1-8.

[7] Yu H, Li H, Mao D, et al. A relationship extraction method for domain knowledge graph construction[J]. World Wide Web, 2020, 23(2): 735-753.

[8] Yang S, Zhu M, Hou J, et al. Deep knowledge tracing with convolutions[J]. arXiv preprint arXiv:2008.01169, 2020.

[9] Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

[10] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

[11] Wang, Benyou, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu and Jakob Grue Simonsen. "On Position Embeddings in BERT." *ICLR* (2021).

[12] Zhong, Zexuan and Danqi Chen. “A Frustratingly Easy Approach for Entity and Relation Extraction.” NAACL (2021).

[13] Goodfellow, Ian J., Yoshua Bengio and Aaron C. Courville. “Deep Learning.” Nature 521 (2015): 436-444.