Computing and Informatics, Vol. 42, 2023, 411–435, doi: 10.31577/cai_2023_2_411

MGCN: MEDICAL RELATION EXTRACTION BASED ON GCN

Yongpan Wang, Yong Liu*, Jianyi Zhang

College of Information Science and Technology
Qingdao University of Science and Technology
Qingdao, 266061, China
e-mail: liuyong@qust.edu.cn

Abstract. With the progress of society and the improvement of living standards, people pay more and more attention to personal health, and WITMED (Wise Information Technology of med) has occupied an important position. The relationship prediction work in the medical field has high requirements on the interpretability of the method, but the relationship between medical entities is complex, and the existing methods are difficult to meet the requirements. This paper proposes a novel medical information relation extraction method MGCN, which combines contextual information to provide global interpretability for relation prediction of medical entities. The method uses Co-occurrence Graph and Graph Convolutional Network to build up a network of relations between entities, uses the Open-world Assumption to construct potential relations between associated entities, and goes through the Knowledge-aware Attention mechanism to give relation prediction for the entity pair of interest. Experiments were conducted on a public medical dataset CTF, MGCN achieved the score of 0.831, demonstrating its effectiveness in medical relation extraction.

Keywords: Relation extraction, co-occurrence graph, attention mechanism, open-world assumption, graph convolutional network

Mathematics Subject Classification 2010: 68U15

* Corresponding author

1 INTRODUCTION

Relation Extraction (RE) is the key task of information extraction, mainly extracting semantic relations between entities from natural language texts, and the results are usually expressed in the form of a triad (subject, relation, object), i.e. (s, r, o). Relation prediction is complementary to this narrow relation extraction, which can combine natural language text information and existing relation triads to reason about the relation between two named entities of interest. It is a complement to the relational network and is also expressed in the form of relational triads. The relation extraction mentioned in this paper is a broad concept, including narrow relation extraction and relation prediction. This type of work is now widely used in knowledge graphs, diagnostic systems, intelligent question and answer, information retrieval, and other related fields.

So far researchers have done a lot of work on relation extraction [1, 2, 3, 4]. At the beginning, people adopted the method based on template matching, using predefined rules or constraints to achieve relationship extraction, the most representative of which is the FASTUS system [5]. With the advancement of technology, traditional machine learning based relation extraction methods have emerged, such as the feature vector based relation extraction method proposed by Miller et al. [6] and the kernel function based relation extraction method proposed by Zelenko et al. [7]. In recent years, deep learning-based relation extraction methods have been proposed by scholars, which have greatly improved the performance of relation extraction. The most classic is LSTM + CRF [8], which is an end-to-end discriminative method. LSTM utilizes past input features, and CRF utilizes sentence-level annotation information, which can effectively utilize past and future annotations to predict current annotations. However, most of the decisions of such neural network methods are black-box operations performed internally, which are difficult to meet the necessary interpretability in the medical field and cannot be directly applied to medical relation extraction tasks.

This paper proposes a medical information relation extraction method MGCN, which can be used in a wide range of medical texts, such as electronic medical records, test reports, medical papers, etc. MGCN uses the co-occurrence graph for modeling to remove sensitive information, which is beneficial to protect patient privacy. The method also utilizes graph convolutional network, which fully incorporate contextual information from medical texts. The open-world assumption is used for relation construction, and the Knowledge-aware Attention mechanism is used to give the final prediction, which provides a reliable basis for the final result. After experiments on dataset, MGCN achieves the F1 score of 0.831, which proves its effectiveness.

The main contributions of this paper are as follows:

1. Aiming at the rigor of the medical field, a highly interpretable method is proposed, which transforms the traditional black-box network computation into relational reasoning.

2. GCN is introduced to compute the relationship between entities, and the information of graph structure data is further utilized while paying attention to the context text information, and the utilization rate of information is improved.
3. The knowledge-aware attention instance encoder is introduced to supervise the reasoning process of the method and it further improves the performance of the method.

2 RELATED WORK

Relation Extraction (RE): RE is to find out the relation between entities in unstructured or semi-structured data, which is part of information extraction and a key step in building a knowledge graph. The existing mainstream relational extraction techniques are classified into three types: supervised learning methods, semi-supervised learning methods, and unsupervised learning methods. The supervised learning methods treat the relation extraction task as a classification problem, design effective features to learn various classification methods based on the training data, and then use the trained classifiers to predict relations. Such methods include many classical methods, such as the DNN [9] proposed by Daojian Zeng, which for the first time equates the relation extraction problem to a relation classification problem and uses deep convolutional neural networks to solve the relation extraction task. The BLSTM [10] proposed by Shu Zhang uses the classical BiLSTM as the main module of the method, reconsiders the lexical feature, and proves its effectiveness. The problem with this type of method is that it requires a large amount of manual annotation of the training corpus, and the corpus annotation work is usually very time-consuming and labor-intensive. The semi-supervised learning methods mainly use bootstrapping for relation extraction. For the relations to be extracted, several seed instances are first set manually, and then the relation template corresponding to the relation and more instances are iteratively extracted from the data. Some representative systems of this type are DIPRE (Dual Iterative Pattern Relation Expansion) [11] proposed by Brin et al. in 1998, NELL (Never-Ending Language Learner) [12] developed by a team led by Professor Tom Mitchell at CMU in 2010, and so on. The semi-supervised learning method of entity relation extraction can partially solve the problem of insufficient number of annotations, but the problem of low accuracy will remain its main challenge for a long time in the future. The unsupervised learning methods assume that pairs of entities with the same semantic relations have similar contextual information. Therefore, we can use the corresponding contextual information of each entity pair to represent the semantic relation of that entity pair, and cluster the semantic relation of all entity pairs. Rozenfeld et al. in 2007 proposed an unsupervised relation identification and extraction system URIES [13], which uses a schema-based contextual representation instead of the context of entity pairs. Yao et al. [14] proposed an unsupervised relation discovery method based on semantic digestion in 2013.

This method uses topic methods to assign entity pairs and their corresponding relation templates to different semantic categories, and then uses clustering methods to map these semantic categories to semantic relations. The effectiveness of such methods depends heavily on how well constraints and heuristics are constructed, and relationships are not as prescriptive as pre-specified relationship types. In comparison, supervised learning methods can extract more effective features with higher accuracy and recall. Therefore, supervised learning methods have received more and more attention from scholars.

In addition, some interesting methods have emerged in the field of relation extraction in the past two years. Xiang Chen proposed an optimization method KnowPrompt [15] based on knowledge co-optimization for text relation extraction (knowledge retrieval, dialogue, question answering) in few-shot scenarios. By learning template words and answer words, knowledge of entities and relations is injected into the methods and their representation is collaboratively optimized under knowledge constraints. Zexuan Zhong proposed a simple and effective end-to-end relation extraction method PURE [16]. The method learns two independent encoders for entity recognition and relation extraction and proposes a new efficient approximation method that achieves large runtime improvements with a small drop in accuracy. Deming Ye proposes a new span representation method PL-Marker [17] that considers the interrelationships between spans (pairs) by strategically wrapping tokens in the encoder. And a neighborhood-oriented packing strategy is proposed to pack the spans with the same starting token into a training instance as much as possible to better distinguish entity boundaries.

Graph Convolutional Network (GCN): Since CNN [18], deep learning methods have achieved high performance for all types of tensors on Euclidean space. However, in addition to the regular data on Euclidean space, there is a large amount of data in the form of topological graphs on non-Euclidean space. A graph data form is shown in Figure 1, which consists of nodes and edges, and nodes connected by edges are neighbors of each other, the number of neighbors of each node is not specified, and there are corresponding signals(information) on each node. Many domain data are represented in this form, such as traffic networks, molecular structures, joint nodes, etc. Traditional convolutional networks are unable to learn such graph-structured data. Based on the need to deal with this topology, the graph convolution method was created. In 2014, Joan Bruna [19] first proposed two different graph convolution construction methods in spatial domain and spectral domain, which laid the foundation for the development of GCN. But its excessive computational complexity and overly large computational parameters limited practical application. Thomas N. Kipf [20] proposed the algorithmic idea of GCN in 2016, and after publishing a related article in 2017, GCN really started to be applied and developed. Then, Michaël Defferrard [21] proposed a second-generation version of the GCN. He has cleverly designed the convolution kernel formula to reduce the number of parameters, re-

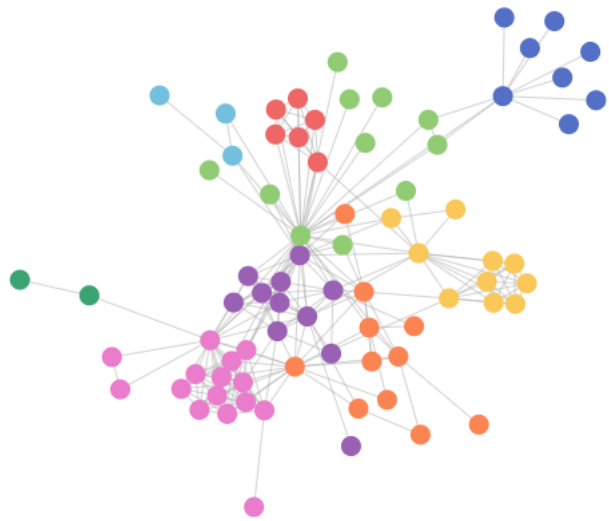


Figure 1. Example of graph data structure

duce the matrix computation, and greatly reduce the computational cost. The rise of GCN has also provided new ideas for solving many natural language processing (NLP) problems. Currently, the way of constructing graph structures by syntactic dependency trees and applying GCN for NLP downstream tasks based on this has been widely used. In addition, there is also work on building graph structures in text through TF-IDF (Term Frequency-Inverse Document Frequency), PMI (Point-wise Mutual Information), sequence relations, lexicon and other information to solve problems using GCN [22, 23, 24, 25]. AGGCN [26] develops a "soft pruning" strategy for the entire dependency tree, transforming the original dependency tree into a fully connected weighted graph. The weights of these graphs are regarded as the correlation strength between nodes and are learned in an end-to-end manner using a self-attention mechanism. At present, there is still a lot of room for GCN to develop in the field of RE.

Open World Assumption (OWA): When making formal descriptions of real-world problems, inevitably the information available is incomplete. For example, we don't know if ibuprofen can cure toothache, but again, this information is indeed useful. A common approach is to use Closed World Assumption (CWA), i.e., if we cannot deduce P or the negation of P in the knowledge base, we add the negation of P to the knowledge base. Another way to deal with incomplete knowledge is to use the Open World Assumption (OWA), which is the opposite of the CWA. OWA is honest about the fact that it does not know the correctness of a proposition that it cannot deduce, with the consequence that the number

of conclusions that can be deduced from the knowledge base is greatly reduced. However, in the semantic Web environment, because of the openness of the Web, the relevant knowledge is likely to be distributed in different places on the Web, so it is inappropriate to use CWA for reasoning on the semantic Web. So, if we want to gather knowledge from different sources in the Semantic Web, we should use OWA. The reasoning in description logic happens to use OWA, so it is indeed suitable as a logical basis for the Semantic Web. In 2016 Ismail Ilkan Ceylan et al. [27] proposed open-world probabilistic databases, as a new probabilistic data method. For unknown facts, this data method assigns any probability value to them from a default probability interval. In 2020 Zhen Wang [28] performed medical entity relation prediction based on corpus-level data and OWA with good results.

3 METHODS

Predicting the relation between entities from a natural language is a very critical task, which can help construct structured knowledge to support a series of downstream tasks such as question answering systems, dialogue systems, inference systems, knowledge graphs, etc. Most of the existing medical information relation extraction methods build deep methods through source texts, and use the attention mechanism to provide local interpretability, which lacks overall global understanding and interpretation. The method MGCN proposed in this paper, for the two medical entities concerned, combines the context information in the medical text and the globality of the medical co-occurrence graph to find their associated entities. Then, the potential relation is constructed using OWA, and finally the final relation prediction is given through the decision module. The overall structure of the method is shown in Figure 2.

3.1 Associated Entity

The first step of the method is to find the associated entities of the entity pair (s, o) of interest. The text information is input into the Bi-LSTM network for word embedding, and then the weights of the relations between nodes are obtained by GCN, as shown in Figure 3. Finally, the top-N nodes that are most closely related to s and o respectively (N is a variable hyperparameter) are found to obtain the set of associated entities.

First, the text information from the library is fed into the Bi-LSTM network to generate word vectors with context, which are then used as the $h^{(0)}$ in the original model. This Bi-LSTM layer is trained jointly with other parts of the network. This has the advantage that the resulting word vector contains both contextual information about word order or disambiguation and provides the correct parse tree on which GCN relies heavily, allowing for more efficient extraction of key information from the sentence. In an L-layer GCN, the input vector of the i^{th} node in the l^{th} layer

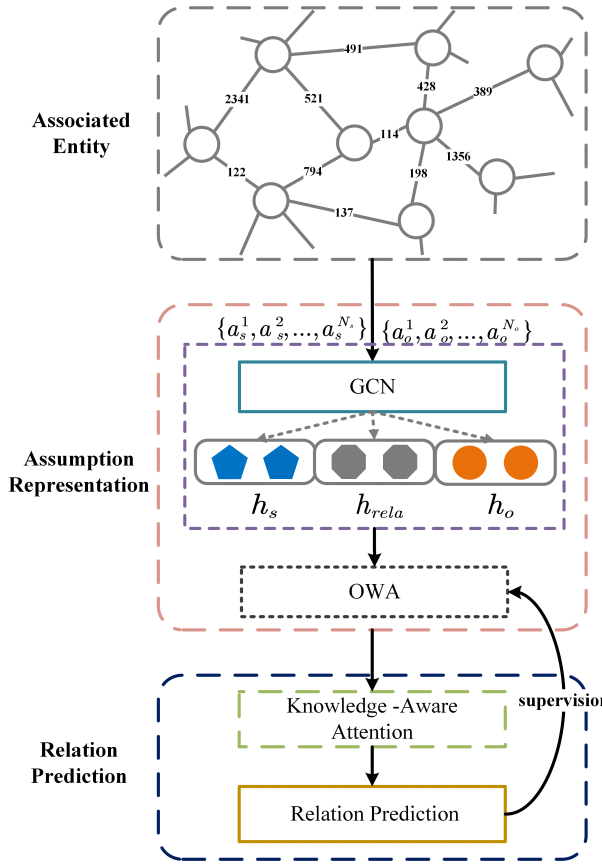


Figure 2. MGCN model overall architecture diagram

is denoted as $h_i^{(l-1)}$ and the output vector is denoted as $h_i^{(l)}$. The graph convolution formula is as follows.

$$h_i^{(l)} = \sigma\left(\sum_{j=1}^{n} A_{ij} W^{(l)} h_j^{(l-1)} + b^{(l)}\right), \tag{1}$$

where $W^{(l)}$ is a linear transformation, $b^{(l)}$ is a bias term, and σ is a nonlinear function (e.g., Relu).

Briefly, during graph convolution, each node collects and aggregates information from neighboring nodes. Convert each dependency tree into an adjacency matrix A and model it uses the graph convolution operation, where $A_{ij} = 1$ if there is dependency edge between nodes i and j. However, because the degree of nodes varies greatly, a direct graph convolution operation in Equation (1) above may lead to very different results for node representation. This may bias the sentence representation

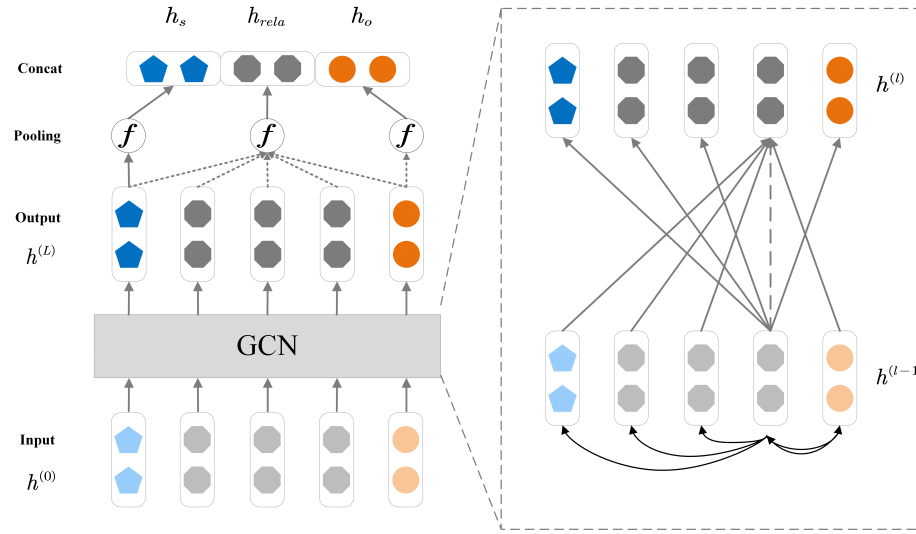


Figure 3. GCN network architecture diagram. The overall architecture is shown on the left, and the detailed calculation method of one-layer graph convolution is shown on the right.

towards nodes with multiple degrees and ignore the information carried by the nodes. In addition, because the nodes in the adjacency matrix have no edge connected to themselves, the information in $h_i^{(l-1)}$ is never passed to $h_i^{(l)}$. To solve these problems, a normalization operation is performed before the data is passed into the nonlinear layer and a self-loop is added to each node in the graph with the following equation:

$$h_i^{(l)} = \sigma\left(\sum_{j=1}^{n} \tilde{A}_{ij} W^{(l)} h_j^{(l-1)} / d_i + b^{(l)}\right), \tag{2}$$

where $\tilde{A} = A + I$, I is the unit matrix of $n \times n$ and $d_i = \sum_{j=1}^{n} \tilde{A}_{ij}$ is the degree of node i in the graph.

This operation is superimposed on the L-layer to obtain a deep GCN network, where $h_1^{(0)}, \ldots, h_n^{(0)}$ is used to represent the input word vectors and $h_1^{(L)}, \ldots, h_n^{(L)}$ to represent the output word vectors. The information transfer between nodes is parallel, and the operations in the network can all be done efficiently by matrix multiplication. After calculating the proximity T of all predicted frames, the confidence of the optimal class of predicted frames is introduced, and the calculation of proximity and confidence is done to describe LT and defined as J. The formula is shown below.

Next, define the model tasks. Let $\mathcal{X} = [x_1, \ldots, x_n]$ denote the sentence, where x_i is the i^{th} word. Identify the subject entity s and the object entity o and correspond them to the two intervals in the sentence: $\mathcal{X}_s = [x_{s_1}, \ldots, x_{s_n}]$ and $\mathcal{X}_o = [x_{o_1}, \ldots, x_{o_n}]$.

Given $\mathcal{X}$, $\mathcal{X}_s$ and $\mathcal{X}_o$, the goal of model is to predict the relation $r \in \mathcal{R}$ ($\mathcal{R}$ is a predefined set of relations) or "no relation" between entities. After applying the L-layer GCN to the word vectors, the implicit representation of each word is obtained, and these representations are directly influenced by their neighbors. In order to use these word representations for relation extraction, the following sentence representations were first obtained (as shown on the left in Figure 3):

$$h_{rela} = f\left(\mathsf{h}^{(L)}\right) = f\left(GCN\left(\mathsf{h}^{(0)}\right)\right), \tag{3}$$

where $\mathsf{h}^{(L)}$ denotes the implicit representation of the overall GCN layer, and $f : \mathbb{R}^{d\times n} \to \mathbb{R}^d$ is the maximum pooling function that maps from the n output vectors to the sentence vectors.

The information close to the entity is usually the core of the relation extraction, and the representation h_s of entity s can be obtained from $\mathsf{h}^{(L)}$, and similarly the representation h_o of entity o can be obtained:

$$h_s = f\left(\mathsf{h}^{(L)}_{s_1:s_2}\right). \tag{4}$$

The final representation for classification is obtained by concatenating the sentence representation and the entity representation and feeding them into a feedforward neural network (FFNN):

$$h_{\text{final}} = \text{FFNN}\left([h_{rela}; h_s; h_o]\right). \tag{5}$$

Then h_{final} is input to the linear layer for Softmax operation to obtain the probability distribution over the relation. The top-N entities are finally selected as associative entities of s/o for subsequent assumption representation.

3.2 Assumption Representation

With associated entities, it is possible to represent assumptions. This method defines the model assumptions as relational interactions between associated entities, as shown in Figure 4. The model can identify (caffeine, may treat, migraine) as a hypothesis, which can help predict that aspirin can treat headache (caffeine and migraine are associated entities of aspirin and headache, respectively). This relational rationale is more specific and easier to understand than the local attention-based explanation strategies widely adopted in NLP. A direct way to obtain this presence relation is to consult the existing medical Knowledge Base (KB), for example (caffeine, may treat, migraine) may be present in SNOMED CT5. This way of obtaining theorems is known as CWA, but in the medical field, the problems of sparsity and incompleteness of KB are serious. Therefore, this method uses OWA to discover more diverse theorems by constructing all potential relations between associated entities.

In OWA, given a pair of entities $e_s, e_o \in \mathcal{V}$, the set of associated entities is defined as $\mathcal{A}(e_s) = \{a_s^i\}_{i=1}^{N_s}$ and $\mathcal{A}(e_o) = \{a_o^j\}_{j=1}^{N_o}$, where N_s, N_o denotes the total number

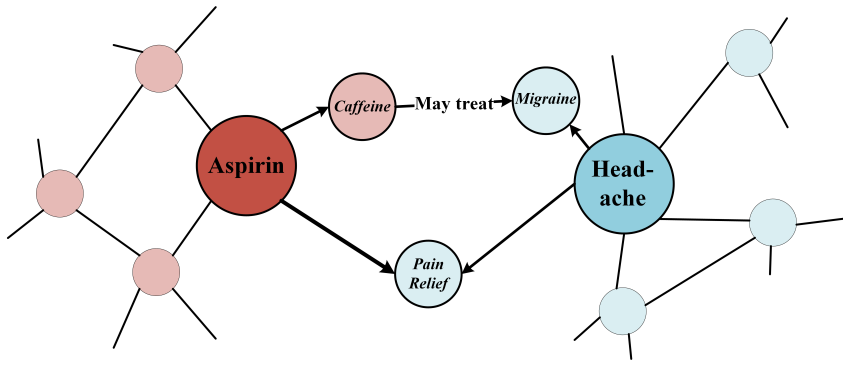


Figure 4. Schematic diagram of potential relation construction

of associated entities. After the previous step, each entity is assigned an embedding vector, which can then be used to measure the probability of maintaining the relation between pairs of associated entities. Given $a_s^i \in \mathcal{A}(e_s), a_o^i \in \mathcal{A}(e_o)$ and relation $r_k \in \mathcal{R}$, define a scoring function to assign a score to the triplet:

$$c_k^{ij} = f\left(a_s^i, r_k, a_o^j\right) = -\left\|h_{a_s^i} + \xi_k - h_{a_o^j}\right\|_1, \tag{6}$$

where $h_{a_s^i}$ and $h_{a_o^j}$ are embedding vectors, the relations are parameterized by a relation matrix $R \in \mathbb{R}^{N_r \times d}$, and ξ_k is a k-level row vector.

Higher scores are obtained when entity pairs and relations are correctly matched. To avoid extremely unreasonable assumptions, the NA relation is defined to represent other irrelevant relations or no relations, and the score is $c_{NA}^{ij} = f\left(a_s^i, \mathrm{NA}, a_o^j\right)$. The OWA principle is expressed by calculating the conditional probability of a relation between a pair of associated entities, the formula is as follows:

$$p\left(r_k | a_s^i, a_o^j\right) = \begin{cases} \frac{\exp(c_k)}{\sum_{s_k \geqslant s_{\mathrm{NA}}} \exp(c_k)}, & c_k > c_{\mathrm{NA}}, \\ 0, & c_k \leqslant c_{\mathrm{NA}}. \end{cases} \tag{7}$$

For each associated entity pair (a_s^i, a_o^j), when the highest value of the relation r is calculated through Equation (7), only the assumption related to r is finally formed. To represent the assumptions, information about all relations for each association pair is integrated into a vector representation, while $p\left(r_k | a_s^i, a_o^j\right)$ is used as the weight of all relations to calculate the assumption representation:

$$\mathsf{a}_{ij} = \rho\left(a_s^i, a_o^j, \mathcal{R}\right) = \sum_{k'=1}^{N_r} p\left(r_{k'} | a_s^i, a_o^j\right) \cdot \xi_{k'}. \tag{8}$$

Combining the entity vector and the relation vector, the final representation of the associated entity's assumption about $\left(a_h^i, a_t^j\right)$ is obtained:

$$e_{ij} = \tanh\left(\left[h_{a_s^i}; h_{a_o^j}; \mathsf{a}_{ij}\right] W_p + b_p\right), \tag{9}$$

where $[\cdot;\cdot]$ denotes the vector connection and $W_p \in \mathbb{R}^{3d\times d_p}$, $b_p \in \mathbb{R}^{d_p}$ are the weight matrix and bias terms of the fully connected network, respectively.

3.3 Relation Prediction

Next comes the relation prediction module which collects all the assumptions, and uses the weighted assumptions information of the target pair to calculate the predicted probability of the relation r. The traditional relation extraction methods only perform relation extraction based on a closed knowledge base, that is, use known factual knowledge for knowledge reasoning. MGCN adopts OWA for relation extraction and uses the calculated probability relation as a given fact to assist the relational reasoning process. Such assumption-based reasoning may lead to certain results that are based entirely on assumptions and are too far from reality. Therefore, we introduce Knowledge-aware Attention to supervise the inference process. The vector v of each instance x of the concerned entity pair is computed using the instance encoder, resulting in a context-based instance representation, which is completely based on known facts. Knowledge-aware Attention will perform attention calculation on assumptions representation and instance representation, to obtain textual relation representation that pays attention to both hypothesis and fact. The introduction of Knowledge-aware Attention will impose certain constraints on assumptions, avoid prediction results that are very inconsistent with facts, and play a supervisory role in the process of assumption reasoning.

This paper designs a new scoring method to measure the confidence of the relation between target entity pairs. Given an entity pair (s, o) and its instance pocket $X_{s,o} = \{x_1, x_2, \ldots, x_m\}$, use the sentence encoder for instance embedding to get $V_{s,o} = \{v_1, v_2, \ldots, v_m\}$. The instance representation thus obtained by the instance encoder is the sentence encoding of each occurrence of the entity of interest in the text. Then use the Knowledge-aware Attention mechanism to get the textual relation representation, which is then used to calculate the relation probability, as shown in Figure 5.

First, the attention weights (similarity or association) between each instance feature vector v_k and assumption representation e_{ij} are calculated:

$$e_k = W_s\left(\tanh\left[v_k : e_{ij}\right]\right) + b_s A_k^i = \frac{\exp\left(c_k\right)}{\sum_{j=1}^{m}\exp\left(c_j\right)}, \tag{10}$$

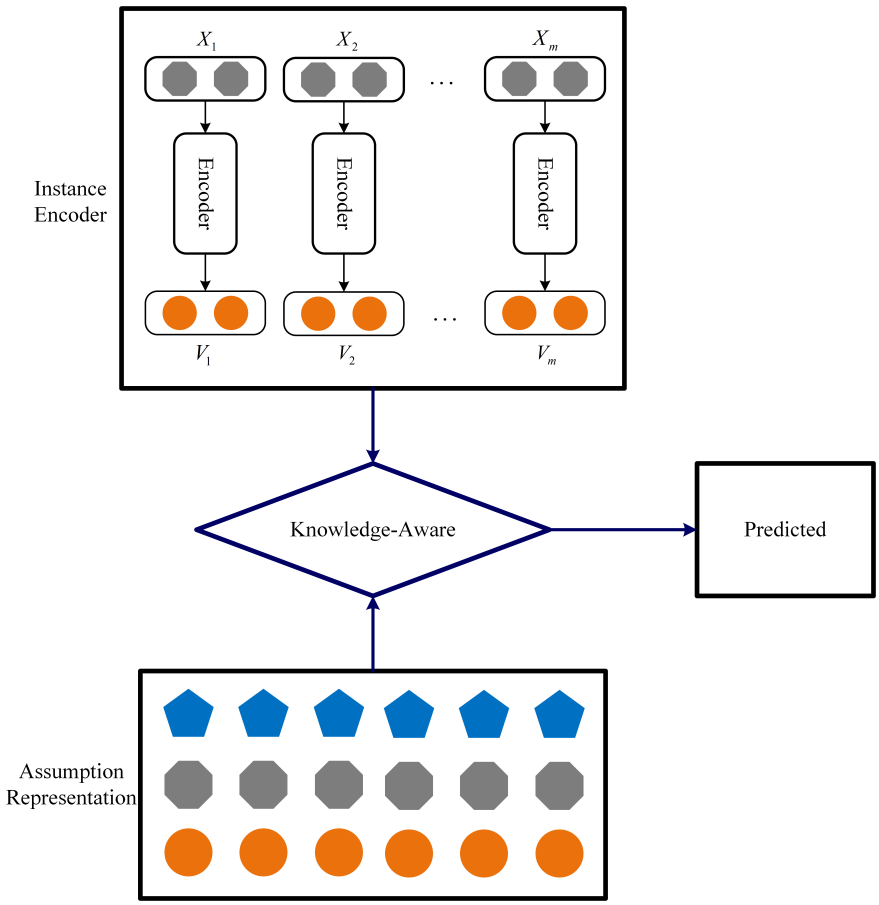


Figure 5. Schematic diagram of the relation prediction stage

where $[x_1 : x_2]$ denotes the vertical connection of x_1 and x_2, W_s is the weight matrix, and b_s is the bias. Then the attention operation is performed on the target entity pair to obtain the corresponding textual relation representation:

$$r^i_{s,o} = ATT\left(e_{ij}, \{v_1, v_2, \ldots, v_m\}\right) g_i = W_g \tanh\left(r_{s,o}\right) \beta_i = \frac{\exp\left(g_i\right)}{\sum_{j=0}^{L-1} \exp\left(g_j\right)}, \tag{11}$$

where W_g is a weight matrix and $r_{s,o}$ is referred to as a query-based function that scores the degree of match between the input textual relation representation and the predicted relation r. The textual relation representation is calculated by:

$$r^i_{s,o} = \beta_i r^i_{s,o}. \tag{12}$$

The textual relation representations of different GCN layers are simply concatenated as the final representation and used to compute the conditional probability $\mathcal{P}(r|s,o)$.

$$r_{s,o} = Concat\left(r_{s,o}^{0}, \ldots, r_{s,o}^{L-1}\right), \tag{13}$$

$$\mathcal{P}(r|s,o) = \frac{\exp(c_r)}{\sum_{\tilde{r}\in R}\exp(\tilde{c}_r)}. \tag{14}$$

At this point, the complete prediction of the relation of the entity pair and their confidence scores are obtained. In addition, to reflect the interpretability of the model, we designed a contribution function O to measure the contribution of all assumptions' representation in the relational inference process:

$$O\left(a_s^i, r_k, a_o^j\right) = \beta_i \times p\left(r_k|a_s^i, a_o^j\right), \tag{15}$$

where β_i is the textual relation representation in Equation (11) and $p(r_k|a_s^i, a_o^j)$ is the relation probability of the associated entity pair in Equation (7).

4 EXPERIMENTS AND RESULTS

This section describes the configuration of the experiments in detail. First, the dataset and evaluation metrics are introduced, and the parameter settings of the experiment and the code running environment are described. Then, MGCN is experimented with a set of comprehensively competitive baseline method on the dataset, and the experimental results are compared and analyzed. Furthermore, to verify the validity of the method rationale, an ablation study was performed.

4.1 Dataset and Evaluation Metrics

Dataset: In order to make full use of the rich resources in the medical field, Finlayson [29] proposed the clinical text frequency (CTF) dataset based on electronic health records (EHR) in 2014. It quantifies pairwise mentions of 3 million terms mapped to 1 million clinical concepts, calculated from the raw text of 20 million clinical records spanning 19 years. The dataset quantifies the correlation between medical entities and eliminates patient privacy information, and its database-level knowledge reserve also provides a reasoning basis for the prediction of medical entity relations. The co-occurrence graph contains 52 804 nodes and 16 197 319 edges, which provides a more concise data form for information researchers in the medical field and greatly promotes the development and utilization of EMR (Electronic Medical Record) resources. After a study of distant supervision of medical texts [30], five medical relations that are more important for clinical decision making were selected. An equal number of negative pairs were extracted by randomly pairing the head and tail entities with the

correct parameter types [31] to help method training. Using the mapping between medical terms and concepts provided by Finlayson et al., relation labels are automatically collected from UMLS (Unified Medical Language System) for training relation prediction. To validate the effectiveness of the method, the dataset was randomly divided into 70 % training, 20 % validation, and 10 % testing in a single experiment.

Med Rela.	Train	Dev	Test
Symptom	14 326	3 001	3 087
May treat	12 924	2 664	2 735
Contraindicates	10 593	2 237	2 197
May prevent	2 113	440	460
Causes	1 389	305	354
Total	41.3 k	8.6 k	8.8 k

Table 1. Dataset statistics

Evaluation Metrics: The evaluation metrics often used for relational extraction tasks are Precision, Recall, and F-Measure. Precision is for the extraction result, which means how many of the samples whose extraction result is the relation R are correct. The TP (True Positive) is the number of correct samples, and the FP (False Positive) is the number of incorrect samples. The formula is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{16}$$

Recall is for the original sample, which indicates how many samples with relation R are correctly extracted. The correct extraction from the sample set with relation R is recorded as TP, and the wrong extraction is recorded as FN (False Negative). Its calculation formula is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{17}$$

For relational extraction, the two metrics, accuracy, and recall, are sometimes contradictory and complementary. In this way, they need to be considered comprehensively. The most common method is the F value, also known as F_{score}. Its calculation formula is:

$$F_{score} = \frac{(\beta^2 + 1) \times \text{Pr} \times \text{Re}}{\text{Pr} + \text{Re}}, \tag{18}$$

where Pr denotes the precision and Re denotes the recall score, β is used to balance the weight of precision and recall in the calculation of F value.

In relation extraction tasks, β is generally taken as 1, and the two metrics are considered equally important. Therefore, the calculation formula of F1 value is:

$$F1 = 2 \times \frac{\mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \tag{19}$$

The models were all evaluated using F1 as the metrics, and the experimental results were averaged over three replicate experiments.

Rela.	UMLS Relations
Symptom of	disease has finding; disease may have found; has associated finding; has manifestation; associated condition of; defining characteristic of
May treat	May treat
Contraindicates	has contraindicated drug
May prevent	may prevent
Causes	cause of; induces; causative agent of
Total	41.3 k

Table 2. Relations correspond to the mapping UMLS semantic relation

4.2 Implementation Details

Experiments adopt the Adam [32] optimization strategies in our method training and use Binary Cross-Entropy (BCE) [33] loss to improve our network performance. For the method to be used on dataset, the initial parameters are set to learning, the rate is 1e−3, batch size is 128. The number of epochs is 200. All training and testing of the methods are implemented on PyTorch 1.7. Repeat the experiment three times and take the average value as the results.

4.3 Comparison Experiments

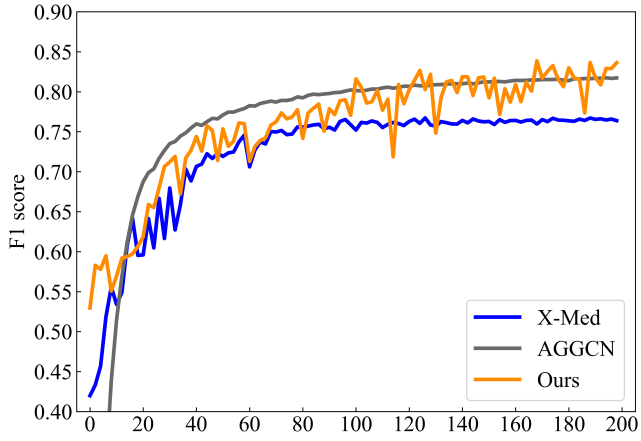
In this section, we compare MGCN with a comprehensive set of relation extraction methods. For predicting the relation between two nodes in a graph, the framework of a neural method usually includes an entity encoder and a relation scoring function. Therefore, different encoders are used to learn entity embeddings and to make a comprehensive comparison. The relation scoring function is unified with RESCAL [34]. The encoders select one word embedding method, Word2vec [35], two graph embedding methods, random-walk based DeepWalk [36], edge-sampling based LINE [37], and one distributional approach REPEL-D [38] for weakly-supervised relation extraction. For graph structure-based relation extraction algorithms, the extended LSTM methods Graph LSTM [39] and bidirectional DAG LSTM [40], attention guided graph convolutional networks (AGGCNs) [26], two newer methods KnowPrompt [15] and PURE [16], and co-occurrence graph-based X-MEDRELA [28] were chosen.

Method	May treat	Symptom	Contrain.	May prevent	Causes	Avg.
Word2vec + RESCAL	0.753	0.764	0.799	0.638	0.650	0.720
DeepWalk + RESCAL	0.701	0.772	0.793	0.623	0.705	0.718
LINE + RESCAL	0.725	0.765	0.800	0.601	0.689	0.716
REPEL-D + RESCAL	0.726	0.769	0.776	0.680	0.707	0.731
Graph LSTM	0.746	0.806	0.743	0.717	0.703	0.743
Bidir DAG LSTM	0.756	0.773	0.769	0.722	0.707	0.745
AGGCN	0.831	0.833	0.801	0.803	0.774	0.828
KnowPrompt	0.836	0.835	0.829	0.814	0.762	0.815
PURE	0.820	0.862	0.833	0.805	0.724	0.809
X-MedRELA	0.805	0.811	0.816	0.676	0.684	0.758
Ours	0.851	0.850	0.832	0.823	0.803	0.831

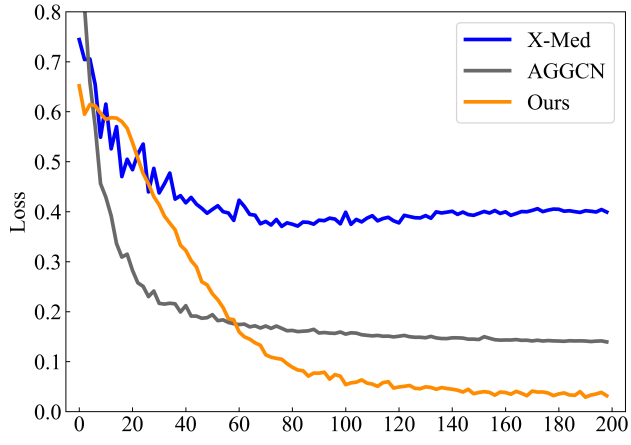
Table 3. Comparison of model predictive performance

Table 3 shows the prediction performance of different methods for F1 scores under each relation prediction task. MGCN obtained a very competitive performance compared to the integrated baseline approach. Specifically, our method achieves substantial improvements in the prediction tasks of "May treat" and "Contraindicates" and performs very competitively in the "Symptom of" and "May prevent" tasks. The poor performance on the "Causes" task may be due to too little training data. This shows that relation extraction based on associations and interactions between entities is effective. Furthermore, compared to those baseline methods that encode graph structures into latent vector representations, MGCN makes full use of co-occurrence graphs, associating context to generate human-understandable rationales. Each stage of our method is interpretable, which can substantially help medical experts.

To demonstrate the effectiveness and convergence of the methods, the F1 and loss curves of X-Med, AGGCN and MGCN at 200 epochs were plotted. As shown in Figure 6 a), the prediction accuracy of all three methods increases rapidly within 40 epochs, and then increases slowly until the best result is achieved at 200 epochs. Among them, the highest F1 value achieved by X-Med is the lowest, and the best result of MGCN is slightly better than that of AGGCN. it can be seen from Figure 6 b) that X-Med and MGCN basically finish converging at 80 epochs, while our method converges approximately at 200 epochs. Comparing the final convergence results, X-Med has the highest loss value at around 0.45; AGGCN also has a poor loss value at around 0.2; while MGCN's loss value has dropped to around 0.04, indicating that the method fits the data well. For deep learning methods, 200 epochs to complete convergence are also a reasonable range. For the performance improvement, the time overhead is worth it. In terms of overall trend, MGCN outperforms X-Med and AGGCN.



a) F1 curve



b) Loss Curve

Figure 6. a) is the FI curve of X-Med vs. MGCN vs. AGGCN and b) is the loss curve of X-Med vs. MGCN vs. AGGCN

4.4 Ablation Study

This section examines the contributions of two main components, namely GCN and Knowledge-aware attention instance encoder. Experiments were conducted on the dataset using the best performing MGCN (w/OWA) method, and the results are shown in Table 4. It can be observed that the introduction of GCN can help the method learn better information aggregation and produce better graph representation, significantly improving the performance of the method. At the same time, adding an attention instance encoder to supervise the inference process of the

method can also further improve the performance. In addition, an ablation study was also carried out for the feedforward layer in the associated entity stage, which confirmed the importance of the feedforward layer in the deep learning method. Without the feedforward layer, the F1 value would drop significantly.

Method	F1
MGCN	0.831
– GCN	0.778
– Attention Instance Encoder (AIE)	0.805
– GCN, AIE	0.758
– Feed-Forward layer(FF)	0.770

Table 4. An ablation study for MGCN model

Performance against Training Data Size. To further test the method performance and explore the effect of different scales of data on the method, a set of experiments were designed. Five training settings (20 %, 40 %, 60 %, 80 %, and 100 % of the training data) were considered in the experiments, and the results are shown in Figure 7. We investigate the performance of MGCN, AGGCN and X-MED on the CTF dataset under different training settings. We investigate the performance of MGCN, AGGCN and X-MED on the CTF dataset under different training settings. At 20 % and 40 % of the training settings, all three methods perform poorly, with MGCN only having a slight advantage, because the performance of deep learning methods relies on large-scale datasets. At 60 % of the training setting, MGCN significantly outperforms AGGCN and X-MED. This means that MGCN has better learning ability when the training data size is average. Under the same amount of training data, MGCN and AGGCN consistently outperform X-MED, and the performance gap becomes more pronounced as the amount of training data increases. When using 100 % training data, the F1 score of MGCN reaches 83.1, which is higher than that of AGGCN of 82.8. These results show that under different scales of data, our method is able to utilize training resources more efficiently and achieve better results.

4.5 Case Studies

This section provides two concrete examples to demonstrate the prediction principles of MGCN to help the reader understand the construction of the method more intuitively.

As shown in Table 5, in order to predict that “cephalosporin” may treat “bacterial infection”, our method will obtain the associated entity “cefuroxime” and “sulbactam” for “cephalosporin”, and the associated entity “viral syndrome’ for “infectious disease” “low grade fever”, “infectious diseases”, and relationships between associated entities. After that, the method will use these five hypothetical principles to predict the relationship between “cephalosporin” and “bacterial infection”,

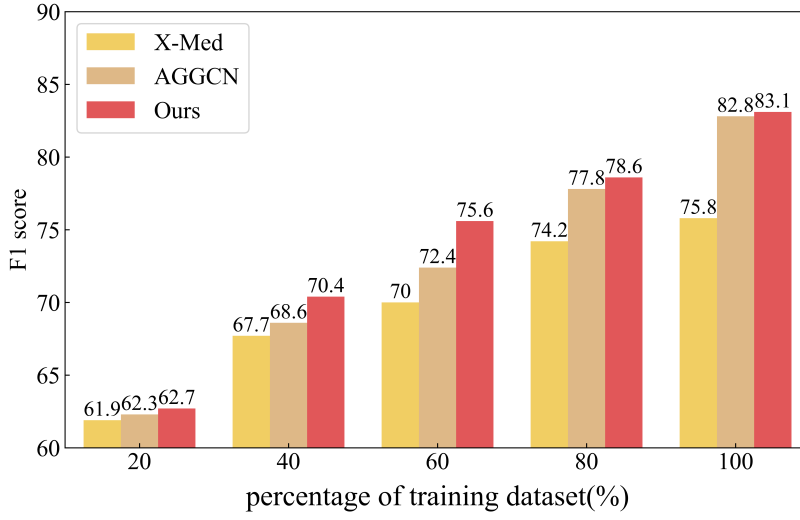


Figure 7. Comparison of MGCN, AGGCN and X-MED against different training data sizes

Subject	Relation	Object
cefuroxime	may treat	viral syndrome
cefuroxime	may treat	low grade fever
cefuroxime	may treat	infectious diseases
cefuroxime	may prevent	low grade fever
sulbactam	may treat	low grade fever
cephalosporins	may treat	bacterial infection

Table 5. Case 1

among which "cefuroxime" may treat "infectious disease" is important to make the final prediction of "possible treatment" theoretical basis. Under the premise of the open-ended hypothesis, doctors may therefore discover new effects of the drug.

Subject	Relation	Object
astepro	may treat	perennial allergic rhinitis
pseudoephedrine	may treat	perennial allergic rhinitis
ciclesonide	may treat	perennial allergic rhinitis
overbite	may treat	perennial allergic rhinitis
diclofenac	may treat	perennial allergic rhinitis
azelastine	may treat	perennial allergic rhinitis

Table 6. Case 2

As shown in Table 6, similarly, the same condition can be treated with different drugs. For the treatment of "perennial allergic rhinitis", the MGCN can give

different medicines (head entities). When one or more of these drugs are known to be effective, doctors can try other drugs to see if they work. Once proven, new drugs can be developed to complement existing drugs. MGCN can make correct predictions based on reasonable principles, providing a theoretical basis to help users understand how the method predictions are performed, and it has an important medical significance.

5 DISCUSSION AND CONCLUSION

Traditional relation extraction methods are all black-box operations, input data sets, output prediction results, and the reasoning process of the method is difficult to visualize. The interpretability of such methods is low and cannot meet the needs of the medical field. For deep learning methods with black-box properties, most of the existing interpretability research use interpretability methods to explain after modeling, such as hidden layer analysis methods, simulation/surrogate methods, sensitivity analysis methods, etc. Different from this kind of research, MGCN itself is an interpretable method. The establishment of the method is based on certain rules, and the decision-making of the method is carried out according to this rule. MGCN performs relation prediction on a given framework, which is set based on the logic of human thinking. For concerned entities, the relevant knowledge is recalled in the first stage, the second stage uses the relevant knowledge to perform relational reasoning, and the third stage gives the prediction result according to the relational reasoning. Under this framework, we can easily understand what each part of the method does and what knowledge is used to make relational predictions. Predicted outcomes for the entity pairs of interest can be traced back to the identified set of associated entities and relational assumptions, as well as the contribution of each assumption to the outcome. In addition to the interpretability of the method itself, each stage of the method can provide a reasoning basis for the results, and has stage interpretability, so MGCN is a method high interpretability. In addition, the method can achieve more accurate and efficient network method tuning and has strong practicability.

This work realizes the relationship extraction of medical information entities and completes the relationship prediction and confidence score of entity pairs in three stages according to different tasks. Unlike existing techniques that rely on multiple different machine learning or deep learning network methods to predict medical entity relationships, MGCN also focuses on the semantic information of the entire corpus while considering OWA. Fully consider the context and spatial structure relationship of the text database, better fit the medical text data characteristics and task characteristics, and finally generate the global optimal prediction theorem. Compared with similar techniques, MGCN is more rational and open in relation prediction, and each stage is interpretable. We believe that MGCN can better assist physicians or practitioners in new medical discoveries and structuring downstream tasks. In the future, this research mainly has the following three exploration direc-

tions. First, consider combining MGCN with state-of-the-art denoising methods to further improve the performance. Secondly, the method is refined on the basis of a small amount of data, so that the method can still achieve better results in the case of scarce data. Finally, MGCN is improved for downstream tasks, enabling it to extract relationships from massive multi-source heterogeneous data and build a medical knowledge graph.

6 DECLARATION OF COMPETING INTEREST

We declare that we have no financial and personal relations with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Acknowledgements

The research was supported by the Center for Ocean Mega-Science, Chinese Academy of Sciences (KEXUE2019GZ04) and GuangHe Fund of Dawning Information Industry Co., Ltd. (GHFUNd202107021586).

REFERENCES

[1] Aone, C.—Halverson, L.—Hampton, T.—Ramos-Santacruz, M.: SRA: Description of the IE2 System Used for MUC-7. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29–May 1, 1998, 1998, https://aclanthology.org/M98-1012.

[2] Zheng, S.—Wang, F.—Bao, H.—Hao, Y.—Zhou, P.—Xu, B.: Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2017, pp. 1227–1236, doi: 10.18653/v1/P17-1113.

[3] Eberts, M.—Ulges, A.: Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training. 2019.

[4] Wei, Z.—Su, J.—Wang, Y.—Tian, Y.—Chang, Y.: A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1476–1488, doi: 10.18653/v1/2020.acl-main.136.

[5] Appelt, D.—Hobbs, J.—Bear, J.—Israel, D.—Kameyama, M.—Martin, D.—Myers, K.—Tyson, M.: SRI International FASTUS System: MUC-6 Test Results and Analysis. 1995, pp. 237–248.

[6] Miller, S.—Fox, H.—Ramshaw, L.—Weischedel, R.: A Novel Use of Statistical Parsing to Extract Information from Text. 2002.

[7] Zelenko, D.—Aone, C.—Richardella, A.: Kernel Methods for Relation Extraction. Journal of Machine Learning Research, Vol. 3, 2003, pp. 1083–1106, doi: 10.3115/1118693.1118703.

[8] Lample, G.—Ballesteros, M.—Subramanian, S.—Kawakami, K.—Dyer, C.: Neural Architectures for Named Entity Recognition. 2016, pp. 260–270, doi: 10.18653/v1/N16-1030.

[9] Zeng, D.—Liu, K.—Lai, S.—Zhou, G.—Zhao, J.: Relation Classification via Convolutional Deep Neural Network. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, 2014, pp. 2335–2344, https://aclanthology.org/C14-1220.

[10] Zhang, S.—Zheng, D.—Hu, X.—Yang, M.: Bidirectional Long Short-Term Memory Networks for Relation Classification. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, 2015, pp. 73–78, https://aclanthology.org/Y15-1009.

[11] Brin, S.: Extracting Patterns and Relations from the World Wide Web. WebDB, 1998.

[12] Carlson, A.—Betteridge, J.—Kisiel, B.—Settles, B.—Hruschka, E.—Mitchell, T.: Toward an Architecture for Never-Ending Language Learning. Vol. 3, 2010.

[13] Rozenfeld, B.—Feldman, R.: High-Performance Unsupervised Relation Extraction from Large Corpora. 2007, pp. 1032–1037, doi: 10.1109/ICDM.2006.82.

[14] Yao, L.—Riedel, S.—Mccallum, A.: Unsupervised Relation Discovery with Sense Disambiguation. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 – Proceedings of the Conference, Vol. 1.

[15] Chen, X.—Zhang, N.—Xie, X.—Deng, S.—Yao, Y.—Tan, C.—Huang, F.—Si, L.—Chen, H.: KnowPrompt: Knowledge-Aware Prompt-Tuning with Synergistic Optimization for Relation Extraction. CoRR, 2021, arXiv: abs/2104.07650.

[16] Zhong, Z.—Chen, D.: A Frustratingly Easy Approach for Entity and Relation Extraction. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 50–61, doi: 10.18653/v1/2021.naacl-main.5.

[17] Ye, D.—Lin, Y.—Sun, M.: Pack Together: Entity and Relation Extraction with Levitated Marker. CoRR, 2021, arXiv: abs/2109.06067.

[18] Liu, C.—Sun, W.—Chao, W.—Che, W.: Convolution Neural Network for Relation Extraction. Advanced Data Mining and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 231–242.

[19] Bruna, J.—Zaremba, W.—Szlam, A.—Lecun, Y.: Spectral Networks and Locally Connected Networks on Graphs. 2013.

[20] Kipf, T.—Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. 2017, arXiv: abs/1609.02907.

[21] Defferrard, M.—Bresson, X.—Vandergheynst, P.: Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. CoRR, 2016, arXiv:

abs/1606.09375.

[22] Liang Yao, C. M.—Luo, Y.: Graph Convolutional Networks for Text Classification. 2019.

[23] Huang, L.—Ma, D.—Li, S.—Zhang, X.—Wang, H.: Text Level Graph Neural Network for Text Classification. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3444–3450, doi: 10.18653/v1/D19-1345.

[24] Gui, T.—Zou, Y.—Zhang, Q.—Peng, M.—Fu, J.—Wei, Z.—Huang, X.: A Lexicon-Based Graph Neural Network for Chinese NER. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 1040–1050, doi: 10.18653/v1/D19-1096.

[25] Sui, D.—Chen, Y.—Liu, K.—Zhao, J.—Liu, S.: Leverage Lexical Knowledge for Chinese Named Entity Recognition via Collaborative Graph Network. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3830–3840, doi: 10.18653/v1/D19-1396.

[26] Guo, Z.—Zhang, Y.—Lu, W.: Attention Guided Graph Convolutional Networks for Relation Extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 241–251, doi: 10.18653/v1/P19-1024.

[27] Ceylan, u. u.—Darwiche, A.—Van den Broeck, G.: Open-World Probabilistic Databases. Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning, AAAI Press, 2016.

[28] Wang, Z.—Lee, J.—Lin, S.—Sun, H.: Rationalizing Medical Relation Prediction from Corpus-Level Statistics. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8078–8092, https://aclanthology.org/2020.acl-main.719.

[29] Finlayson, S.—LePendu, P.—Shah, N.: Building the Graph of Medicine from Millions of Clinical Narratives. Scientific Data, Vol. 1, 2014, doi: 10.1038/sdata.2014.32.

[30] Wang, C.—Fan, J.: Medical Relation Extraction with Manifold Models. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers), Association for Computational Linguistics, 2014, pp. 828–838, doi: 10.3115/v1/P14-1078.

[31] Wang, C.—Cao, L.—Fan, J.: Building Joint Spaces for Relation Extraction. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, IJCAI'16, 2016, pp. 2936–2942.

[32] Kingma, D. P.—Ba, J.: Adam: A Method for Stochastic Optimization. 2014.

[33] Boer, P. T.—Kroese, D.—Mannor, S.—Rubinstein, R.: A Tutorial on the Cross-Entropy Method. Annals of Operations Research, Vol. 134, 2005, pp. 19–67,

doi: 10.1007/s10479-005-5724-z.

[34] Nickel, M.—Tresp, V.—Kriegel, H. P.: A Three-Way Model for Collective Learning on Multi-Relational Data. Proceedings of the 28th International Conference on International Conference on Machine Learning, Omnipress, Madison, WI, USA, ICML'11, 2011, pp. 809–816.

[35] Mikolov, T.—Sutskever, I.—Chen, K.—Corrado, G.—Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2, Curran Associates Inc., Red Hook, NY, USA, NIPS'13, 2013, pp. 3111–3119.

[36] Perozzi, B.—Al-Rfou, R.—Skiena, S.: Deepwalk: Online Learning of Social Representations. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, doi: 10.1145/2623330.2623732.

[37] Tang, J.—Qu, M.—Wang, M.—Zhang, M.—Yan, J.—Mei, Q.: LINE: Large-Scale Information Network Embedding. Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2015, doi: 10.1145/2736277.2741093.

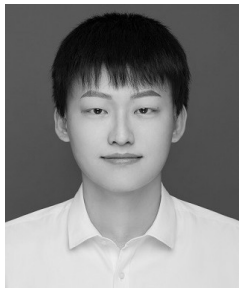
[38] Qu, M.—Ren, X.—Zhang, Y.—Han, J.: Weakly-Supervised Relation Extraction by Pattern-Enhanced Embedding Learning. Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, doi: 10.1145/3178876.3186024.

[39] Peng, N.—Poon, H.—Quirk, C.—Toutanova, K.—Yih, W. t.: Cross-Sentence N-Ary Relation Extraction with Graph LSTMs. Transactions of the Association for Computational Linguistics, Vol. 5, 2017, pp. 101–115, doi: 10.1162/tacl_a_00049.

[40] Song, L.—Zhang, Y.—Wang, Z.—Gildea, D.: N-Ary Relation Extraction Using Graph-State LSTM. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 2226–2235, doi: 10.18653/v1/D18-1246.

Yongpan Wang is currently pursuing her Master's degree at the School of Information Science and Technology, Qingdao University of Science and Technology. Her research interests include information extraction and knowledge graphs.

Yong Liu graduated from the Ocean University of China in 2011, majoring in computer application technology. She is dedicated to the research of intelligent identification and quantitative analysis of marine organisms, medical knowledge graph and intelligent medical big data. So far, she has obtained one national invention patent, seven software copyrights, and published more than 20 research papers (SCI/EI). She has published two textbooks.



Jianyi Zhang is currently pursuing his Master's degree at the School of Information Science and Technology, Qingdao University of Science and Technology. His research interests include object detection, medical image segmentation, and natural language processing.