IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 20, 2020, accepted November 5, 2020, date of publication November 11, 2020,
date of current version November 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037333

Chinese Short Text Entity Disambiguation Based on the Dual-Channel Hybrid Network

LITING JIANG[1,2,4], GULILA ALTENBEK[1], DI WU[2], YAJING MA[1,2,4], AND HAYINAER AIERZHATI[3]

[1]College of Information Science and Engineering, Xinjiang University, Ürümqi 830046, China
[2]Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Ürümqi 830046, China
[3]School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China
[4]National Language Resource Monitoring and Research Center on Minority Languages, Ürümqi 830046, China

Corresponding author: Gulila Altenbek (gla@xju.edu.cn)

This work was supported in part by the Science Fund Project of Xinjiang University under Grant BS180250, and in part by the National Natural Science Foundation of China under Grant 61363062 and Grant 62062062.

ABSTRACT Entity disambiguation refers to the accurate inference of the real mention of an entity with the same name according to the context. Most existing studies focused on long texts, for short texts, the performance has been unsatisfactory due to sparsity. In this paper, we treat the entity disambiguation task as a classification problem. we propose a novel neural network-based capsule network and convolutional neural network for entity disambiguation, leveraging full semantic information of short text data. In particular, a self-attention mechanism is utilized to further filter the semantic information extracted from the capsule network. On the other hand, a convolutional neural network with combined pooling is established to capture semantics from another channel. In the end, the semantic features obtained by the above models are combined through a fully connected layer to complete the task of entity disambiguation. The experimental results on the CCKS 2019 entity linking dataset showed that the dual-channel hybrid network proposed in this paper achieved an F1-score of 88.04%, which is superior to that of the existing mainstream deep learning model, thereby verifying the effectiveness of the model.

INDEX TERMS Entity disambiguation, short text, convolutional neural networks, capsule networks.

I. INTRODUCTION

Due to the current information explosion, to make better use of massive data, researchers have launched in-depth research on knowledge graphs. Entity disambiguation is an important task in knowledge graphs and a key technology in the field of information extraction and integration that affects the accuracy of many downstream tasks in natural language processing, such as information retrieval and intelligent question answering. Recently, the research on entity disambiguation based on knowledge graphs has gradually increased, for example, Moon *et al.* [1], Agarwal *et al.* [2]. Causes of entity ambiguity can be classified into diversity and ambiguity, namely, synonymy and polysemy [3]. The aim of entity disambiguation is to solve the ambiguity of entities with the same name. For instance, ''Nike created the Air Jordan brand for basketball superstar Jordan''. The two Jordans have different references in this sentence. Scholars expect to obtain accurate and unambiguous information through entity disambiguation, but how to efficiently obtain the context information and complete the real entity reference remains a challenging problem. Compared with long text, the short text is shorter in length, contains very little effective information, and has sparse semantic expression. Due to these characteristics, the model will not learn text features well, which makes disambiguation accuracy difficult to improve. In this paper, our research focus is to fully extract semantics from limited texts and complete the task of entity disambiguation. Most of the existing schemes are through showing or implicit text representation. Showing text representation follows traditional NLP steps, including chunking, tagging and syntax analysis. Implicit text representation uses neural language model (NLM) representation and maps text to semantic space for semantic expansion [4]. In this paper, the semantic information is expanded by expanding the text length.

To better solve the difficulty of Chinese short text in entity disambiguation, this paper proposes a dual-channel hybrid

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Shariq Imran.

network model that combines the different features generated by the two models. First, the candidate entity set is obtained from the external knowledge base. Then, the text of the mention and the candidate entity set are spliced one by one to expand the text length so that it has sufficient context information. The unique identification of the entity is used to make the classification label at the same time. Last, the text is input into the BERT model for feature extraction, and the trained features are taken as the input of the CNN model with max-pooling and mean-pooling to capture semantic features. A capsule network is used to capture the semantic features generated by BERT, and an attention mechanism is used to extract the important information of capsule network features. The abovementioned learned semantic knowledge is combined, and the classification is completed by the fully connected layer.

The main contributions of the paper are as follows:

(1) This paper counts the mentions in the training data that cannot link to the knowledge base, places the mentions with the same entity name in a list, finds the mentions with the same entity name in the external knowledge base, and uses the unique identification of the mentions to splice the text of the mention with the text of the knowledge base to construct the text that can be trained by the model.

(2) This paper proposes a DHCN model, which is a combination of a pooling CNN and capsule network with attention. The semantic information generated by the two models is fused in parallel to complete the disambiguation. To the best of our knowledge, this is the first time that capsules have been used in Chinese entity disambiguation tasks.

(3) In this paper, the proposed model achieves a state-of-the-art result in the task of disambiguation of Chinese short text entities.

II. RELATED WORK

Entity disambiguation tasks can be divided into two types: entity disambiguation for structured text and entity disambiguation for unstructured text. At present, most disambiguation methods are based on entity linking technology, which mainly involves methods based on the following models: probability generation models, topic models, methods based on graphs, and method based on deep neural network. The essence of these methods is to calculate the semantic similarity between entity mention and candidate entities and select the candidate entity with the largest similarity score as the target entity [5], such as the traditional context similarity [6].

In the past, Kataria *et al.* [7] proposed a hierarchical topic model for entity disambiguation. This model is a semi-supervised hierarchical model that can be learned in semantic space. Liu *et al.* [8] proposed a disambiguation algorithm using a standard space model and HAC clustering algorithm. Bagga and Baldwin [9] used a vector space model (VSM) to calculate the similarity between entity mention vectors. Bollegala *et al.* [10] first used a specific name to identify each document in the set and then clustered the set of documents. However, clustering-based methods have high time complexity and typically use semantic similarity; the use of a different similarity calculation method would directly affect the clustering effect.

Due to the powerful nonlinear fitting capabilities of neural networks, deep learning models are widely used in entity disambiguation. For example, He *et al.* [11] applied a deep learning model to entity disambiguation for the first time. He directly optimized the document and entity representation for a given similarity measure, used a Stacked Denoise AutoEncoder to learn the initial document representation in the unsupervised pre-training stage, and finally performed supervised fine-tuning. However, due to the sparse semantics of short texts, deep neural networks are needed to further explore the semantics, but excessively deep-stacked automatic encoders are prone to the vanishing gradient problem. Shengze *et al.* [12] proposed the SA-ESF algorithm, which use the stacked Bi-LSTM neural network with a dual attention mechanism to calculate the correlation between entities from the aspects of the entity and its context features, entity description features, etc. Finally, the final disambiguation result is obtained using the similarity score sum and prior probability between each mention and its candidate entities. However, the proposed model focuses on the long-distance dependence of sentences and does not consider the local features of texts, while disambiguation mainly depends on a few words around the entities. Gupta *et al.* [13] captured lexical, syntactic, and local text information of the mention according to a bidirectional LSTM encoder and used a CNN to model entity documents in combination with fine-grained types of structured information sources. Structured information includes a description of the knowledge base. However, the quality of the knowledge base directly impacts the result of entity disambiguation, and this paper does not consider the situation in which the entities in the data set do not match those in the knowledge base. Phan *et al.* [14] proposed paired connections for collective entity disambiguation. A pair-linking algorithm was used to approximate the solution of MINTREE (tree-based entity disambiguation target) by simulating the Kruskal algorithm, thus obtaining the correct entity set. However, this method uses the skip-gram method in Word2vec to complete text vectorization representation. The word vector generated by this method corresponds one-to-one to the word itself and cannot reflect the true meaning of the same word in different contexts. Huang *et al.* [15] used a joint ranking framework to find similar or related entities to eliminate ambiguity. They proposed extending the conceptual short text embedding model with an entity disambiguation framework and using the attention model to select relevant words for prediction. When addressing short text NLP tasks, relatively little useful information is contained in the text, and the attention mechanism alone is unable to acquire the full semantic knowledge.

Many models have been proposed by scholars to capture the lexical, syntactic and conceptual information of the text to complete the task of entity disambiguation. Semantic features are extracted through models, as shown in the experimental

results, the accuracy obtained through these models can be further improved.

Capsule networks were originally applied to image tasks and have since been widely used in the field of natural language processing. Yang *et al.* [16], [17] and Zhao [18] proposed a text classification model based on a variant of CapsNets. The author explored two capsule network architectures, namely, Capsule-A and Capsule-B. Capsule-A is similar to the capsule network used in dynamic routing. Capsule-B uses three convolution kernels of different sizes to run convolution in parallel in the n-gram convolutional layer to extract more comprehensive semantic information. Yang *et al.* [16] proposed three strategies to stabilize the dynamic routing process to reduce the interference of noise capsules that may contain "background" information or that have not been successfully trained. Experiments prove the effectiveness of the model.

Although the above deep learning algorithms have achieved good results, some problems remain. For short text, the features provided by the context are limited, and the features cannot be well represented, which leads to some models not fully learning the features of the text. Therefore, some scholars proposed pre-training a model to address this problem [19], connecting the text and description as the input of BERT, and then classifying the vector output by [CLS] position together with the start position and end position vectors of the candidate entities. However, because the features extracted by the BERT model are relatively broad and noisy, we believe performance on this task can be further improved. In this paper, a hybrid pooled CNN model and a capsule network with an attention mechanism are applied to extract semantic information obtained by BERT. The obtained semantic information is fused in a parallel manner, which avoids the vanishing gradient problem that may occur when the depth of the learning model is excessive. The advantages of the two models are fully applied, and the results are better than those obtained by the existing mainstream models.

III. THE PROPOSED APPROACH

Given a sentence x = (x1, x2, x3,. . . , xn), it is vectorized and represented as the input of the model. The model matches each candidate entity in the external knowledge base according to the context information of the entity to judge what type an entity in the short text belongs to. For example, 苹果系列的电子产品性能都很好 (translation: The performance of Apple series electronic products is very good), in which ''苹果'' (translation: Apple) stands for the company Apple Inc, the category of mobile phones in this sentence, not the fruit apple. According to the learned model, judge which category ''苹果'' (translation: iPhone) belongs to Apple or iPhone.

This paper proposes a dual-channel hybrid network model to further improve the accuracy of Chinese entity disambiguation. First, the BERT model is used to extract the features of the pre-processed data, and the obtained semantic vector representation is used as the input of the CNN model and capsule network. The features extracted from the model are fused, and the fused information is classified by the fully connected layer. By means of this approach, the semantic features of Chinese short texts are enriched, and the accuracy of entity disambiguation is improved. The model diagram is shown in Figure1.

The left part of the figure shows the features of the BERT model for text extraction. The CNN model is shown in the lower half of the figure, and the capsule network and attention mechanism are shown in the upper half of the figure. The ReLU activation function is used in all connection layers. The right part of the figure is the model fusion stage. In this paper, the Add function is used to carry out the semantic fusion of the two sub-models.

A. VECTORIZED TEXT REPRESENTATION

Natural language cannot be directly numerically calculated by neural networks, the text should be vectorized. Text vectorization [20]–[23] represents text as a series of vectors that can express text semantics. Such approaches to text representation and can be roughly divided into two categories: traditional language model and structure or serialization model. Traditional representation methods include one-hot representation, which was initially widely used in the field of natural language processing. Since the codes generated by the method cannot reflect the correlation of words, scholars have developed distributed representation models composed of neural networks, encoders, etc. This unsupervised method is based on the distribution hypothesis (words appearing in the same context often have similar meanings). The most commonly used models are Word2vec [24] and GloVe [25], which represent a word as a vector with unified meaning. The difference is that GloVe considers the vocabulary co-occurrence matrix when training word vectors. However, due to the one-to-one relationship between words and vectors, the problem of polysemy cannot be solved by this approach, and specific tasks cannot be dynamically optimized. Compared with the above two methods, Google's BERT model [26] can effectively reflect the real semantics of words in sentences. Its internal structure is composed of the encoder part of the bi-directional transformer [27], and the semantic vectors generated by the model can accurately reflect the semantic information of words in the text. The BERT model, which is widely used in various NLP tasks [28]–[30], is shown in Figure.2

Due to the limited context provided by short text, to make the model learn better, this paper splices the entity text sequence to be disambiguated with the corresponding entity text sequence in the external knowledge base to create the input of the BERT model. The sequence is as follows:

$$s = (x_1, x_2, \ldots, x_n, c_1, c_2, \ldots, c_m). \tag{1}$$

where x_1 x_n is the entity text to be disambiguated, c_1 c_m is the candidate entity text, and s is the spliced text.

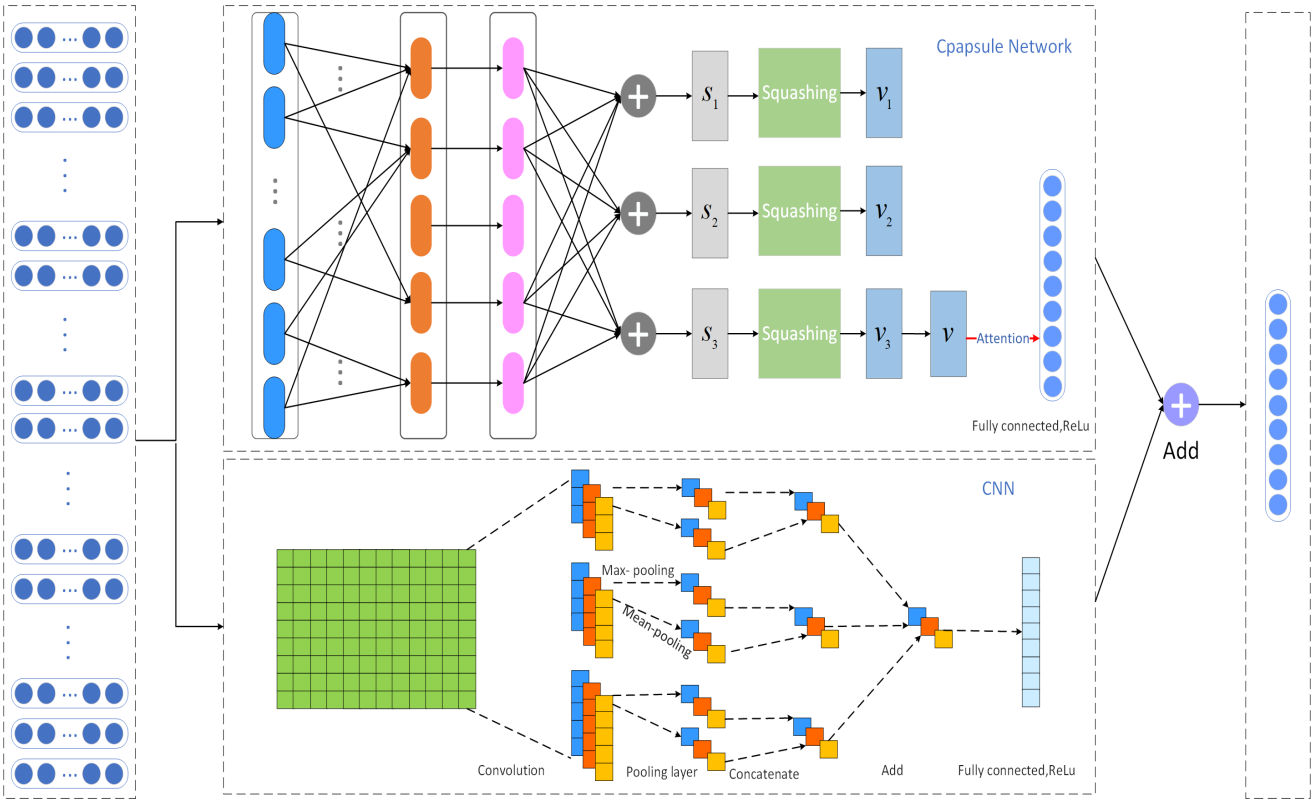

FIGURE 1. Dual-channel hybrid network model.

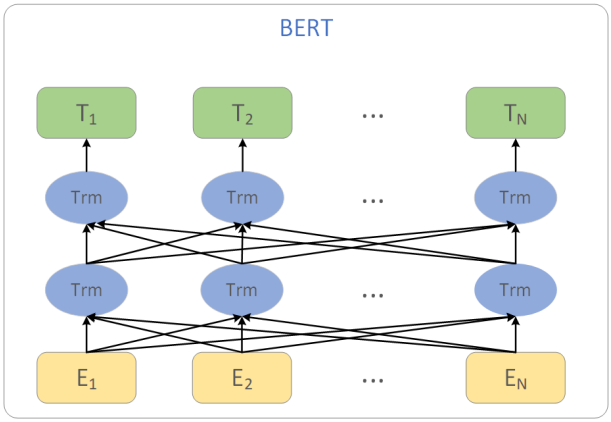

FIGURE 2. Structure of BERT model.

B. CONVOLUTIONAL NEURAL NETWORK

Kim [31] improved the convolutional neural network (CNN) to complete the task of text classification, and this method has been widely used in various downstream NLP tasks. The model has the characteristics of local connection and weight sharing [32], [33]. The CNN is generally composed of three parts, namely, the convolutional layer, pool layer, and fully connected layer. The function of the convolutional layer is to extract features. Local word order information from the input long text sequence is used to complete the extracted features. In this paper, windows with convolution kernel sizes of 3, 4, and 5 are used for convolution to obtain local features. The formula as follows.

$$y_i = f(W_k \cdot H_{i:i+k+1} + b), \quad k = 3, 4, 5. \tag{2}$$

$$Y = y_1 \oplus y_2 \oplus \ldots \oplus y_{n-k+1}. \tag{3}$$

where b represents the bias, W_k represents the weight matrix corresponding to different convolution kernels, i is the i-th feature value, k is the size of the convolution kernel in the convolution operation, f is the activation function, y_i represents the output result after convolution, the operation $\oplus$ is a vector, y is the final feature obtained.

The function of the pooling layer is to filter noise and extract meaningful semantic features. There are two common pooling methods, max-pooling and mean-pooling: max-pooling effectively extracts the most significant semantic information, while mean-pooling considers all semantic information. This paper integrates the semantic information output by the above two pooling operations and takes into account the advantages of both approaches. The formula is shown as follows.

$$P_{max} = MaxPooling(Y). \tag{4}$$

$$P_{mean} = MeanPooling(Y). \tag{5}$$

$$P = P_{Max} \oplus P_{Mean}. \tag{6}$$

where P_{max} represents the result of max-pooling, P_{mean} represents the result of mean-pooling, and P represents the result of the fusion of the two pooling methods.

C. CAPSULE NETWORK AND ATTENTION MECHANISM

The capsule network is a complex neural network proposed by Sabour *et al.* [34] and Hinton *et al.* [35] that is used in image processing tasks. The network was later adapted for use in NLP tasks [36], [37]. The model is composed of a convolutional layer and a dynamic routing layer. The core of the dynamic routing layer is the dynamic routing algorithm, which uses a vector representation of capsule nodes instead of the scalar value representation of neuron nodes used in traditional neural networks to replace the pooling algorithm, thus extracting more abundant location space information instead of the scalar generated by the pooling operation. To capture the semantic information of Chinese short texts more fully, this paper completes the learning of semantic dependency via a two-layer capsule network. The formula for dynamic routing is as follows.

$$u^a = W^a v^a. \tag{7}$$
$$u^b = W^b v^b. \tag{8}$$
$$b_a^0 = 0, b_b^0 = 0. \tag{9}$$
$$c_a^r, c_b^r = softmax(b_a^{r-1}, b_b^{r-1}). \tag{10}$$
$$s^r = c_a u^a + c_b u_b \tag{11}$$
$$n^r = \frac{\|s^r\|^2}{1+\|s^r\|^2}\frac{s^r}{\|s^r\|^2}. \tag{12}$$
$$b_i^r = b_i^{r-1} + n^r \cdot u^i \tag{13}$$

where u^a and u^b are the outputs of the upper layer capsule, W^a and W^b are the linear transformation matrices, b_a^0 and b_b^0 are the initialization coupling coefficients before the dynamic routing calculation, r represents the number of dynamic routes, and c_a and c_b are the coupling coefficients in the dynamic routing process. Formulas (10),(11),(12), and (13) represent the calculation process of dynamic routing, where formula (10) ensures that the sum of the coupling coefficients is 1. Formula (11) calculates the output vector of the capsule network obtained after linear transformation according to the coupling coefficients and capsule output of the previous layer, formula (12) is a squashing operation that changes s^r into a vector of length 1 without changing its direction n^r, and Formula (13) is used to update the coupling coefficient. The final target vector is obtained through the above four formulas.

Bahdanau *et al.* [38] introduced an attention mechanism into NLP tasks for the first time. The core idea is that not all words in a sentence have the same contribution to NLP tasks. In this paper, the attention mechanism is used to further calculate the extracted semantic features, and its formulas are shown in formulas (14), (15) and (16).

$$u_{it} = \tanh(W_w n_t^r + b_w). \tag{14}$$
$$a_{it} = \frac{exp(u_{it}^T u_w)}{\sum_t exp(u_{it}^T u_w)}. \tag{15}$$
$$s_i = \sum_t a_{it} n_t^r. \tag{16}$$

where n_t^r represents the output of the capsule network, W_w represents the weight matrix, b_w represents the offset, u_{it} represents the similarity, a_{it} represents the weight, and s_i represents the final result. The semantic similarity is calculated by formula (14), the weight coefficient of each semantic feature is calculated by formula (15), and the semantic information is weighted and summed by formula (16) [39].

The traditional capsule network flattens the semantic information output via the dynamic routing layer and finally outputs the final result via the fully connected layer. To make the model better able to learn the limited semantic information in short texts, this paper takes the feature output by the capsule network as the input of the self-attention mechanism layer, and the self-attention mechanism conducts further semantic learning.

D. FUSION LAYER

By means of the parallel structure of the CNN model and capsule network + attention, this paper fuses the extracted features via addition, as shown in formula (17).

$$A^c = Add(O_{cnn}, O_{Capsules}). \tag{17}$$

O_{cnn} is the feature extracted by the CNN, $O_{Capsules}$ s the feature extracted by the capsule network, and A^c is the fused feature from the CNN and capsule network.

Finally, the fused semantic features are input into the fully connected layer with a sigmoid activation function for classification, as shown in formulas (18) and (19).

$$f(z) = \frac{1}{1+exp(-z)}. \tag{18}$$
$$y = sigmoid(W_a A^c + b). \tag{19}$$

where z is the output after feature fusion. This approach uses a fully connected weight matrix and a combined feature matrix.

E. CROSS-ENTROPY FUNCTION

The gradient descent algorithm Adam model is used for training, and the objective function is minimized [40]. At the same time, the minimum cross-entropy is used to optimize the model, and the early stopping method and cross-validation method in linear regularization are used to solve the over-fitting problem. The cross-entropy function is shown in formula (20).

$$loss = \frac{1}{N}\sum_{n=1}^{N} \hat{y}^{(i)} \log y^{(i)} + (1-\hat{y}^{(i)}) \log(1-y^{(i)}) \tag{20}$$

where *loss* is the loss function, N is the input sample, $\hat{y}^{(i)}$ is the actual category of the i-th sample, and $y^{(i)}$ is the model prediction category.

TABLE 1. Sample experimental data.

id	text
1	"text_id": "24219", "text": "苹果最新电脑《the new macbook》-广告-高清", "mention_data": ["kb_id": "218104", "mention": "苹果", "offset": "0", "kb_id": "405868", "mention": "电脑", "offset": "4", "kb_id": "69522", "mention": "广告", "offset": "24"]
2	"text_id": "82741", "text": "我们都爱笑。。到底是个什么节目 完全没有笑点啊", "mention_data": ["kb_id": "291480", "mention": "节目", "offset": "13"]
3	"text_id": "33203", "text": "河北经济日报·数字报", "mention_data": ["kb_id": "285372", "mention": "河北经济日报", "offset": "0", "kb_id": "222737", "mention": "数字报", "offset": "7"]

TABLE 2. Experimental parameter setting.

parameter	The parameter value
learn rate	0.0001
batch-size	64
dropout	0.15
epoch	10
Filters	[3,4,5]
num_capsules	3
routing	3

F. ALGORITHM DESCRIPTION

We obtain the candidate entity set from the knowledge base according to the entity to be disambiguated, splice the entities one by one to construct the classified text, use the CNN model with the combined pooling and capsule network with attention mechanism to extract the features, and train and predict the text after the features are fused. The specific process is shown in the algorithm. Dual-channel hybrid network model is shown in Algorithm.1:

IV. EXPERIMENT

A. DATA SET

The data set is obtained from the entity linking task for Chinese short text from the China Conference on Knowledge Graph and Semantic Computing (CCKS 2019).[1] There are 90,000 pieces of data in the data set and 399,252 pieces of data in the external knowledge base, all of which are short Chinese texts. The data came from microblogs, user dialogue content, article titles, etc. In this paper, 8000 pieces of data were randomly selected as the training sets and 1000 pieces of data were selected as the test set. Samples of experimental data are shown in Table1.

B. PARAMETER SETTINGS

The experimental parameters for the dual-channel hybrid model are shown in Table.2

C. EVALUATION CRITERIA

To measure the performance of the model in entity disambiguation tasks, this paper uses precision (P), recall (R), and F1-Score[41] as evaluation indexes. The specific calculations

[1]http://www.ccks2019.cn/

TABLE 3. Confusion matrix.

	Positive	Negative
True	True Positive(TP)	True Negative(TN)
False	False Positive(FP)	False Negative(FN)

are shown in formula (21), (22) and (23).

$$P = \frac{TP}{TP + FP}. \tag{21}$$

$$R = \frac{TP}{TP + FN}. \tag{22}$$

$$F1 = \frac{2 * P * R}{P + R}. \tag{23}$$

where the meanings of TP, TN, FP and FN are shown in the confusion matrix of Table.3

D. EXPERIMENTAL RESULTS AND ANALYSIS

1) INTRODUCTION TO COMPARE MODEL

To verify the effectiveness of the dual-channel hybrid network model, this paper selects several deep learning models for comparison.

- Word2vec(C+W)-Bi-LSTM: Bi-LSTM is a variant of an RNN that can better capture context information in sentences. Char-level, and word-level combination representation vectors are used as inputs to the Bi-LSTM model.
- Word2vec(C+W+P)-Bi-LSTM: Char-level features and position features based on word vectors are added to enrich the text features. The multi-feature fusion result is used as the input to the Bi-LSTM model to obtain the classification results.
- BERT-Bi-LSTM: The output of BERT is input into the Bi-LSTM model. The semantic information extracted from BERT is further studied through the input gate, forget gate and output gate of the model.
- BERT-CNN: A CNN model is used to learn the semantic information extracted from the BERT model to complete the disambiguation task.
- BERT-CNN-ComPooling: After using the CNN model to extract local semantic features of the text, mean-pooling and max-pooling are used to extract features. The features are then fused and classified.
- BERT-Capsule: A capsule network is used to extract word vectors generated by BERT, thus generating more comprehensive semantic information.
- BERT-Capsule-att: A self-attention mechanism is used to obtain more important information generated by the capsule network.
- BERT-DCHN: The CNN model is combined with a capsule network in parallel. The capsule network solves the shortcomings of the CNN model and integrates the two features after introducing a self-attention mechanism.

Algorithm 1 Dual-Channel Hybrid Network Model

Input: Entity to be disambiguated, knowledge base;
Output: 0 or 1;

Begin
The data is preprocessed, the special char in the data are converted, and the book titles are deleted, etc. The data set is divided into training set and test set;
Obtaining a set of candidate entities from the knowledge base, splicing the entities to be disambiguated with the candidate entities, and constructing classification labels according to kb-id and subject-id;
kfold =KFold(splits=5,shuffle=False)
For i,(traindex, valindex)∈fold(train[0])
Beign
    The training text is input into the BERT model to obtain a representation vector;Use the representation vector as the input to the capsule Networks and CNN models;
    concatx=[],concaty=[];
    For index,filtersize∈[3,4,5]
        Beign
        Convolution layer,convolution kernel size [3,4,5];
        mean-pooling,append to concatx;
        max-pooling,append to concaty;
        End
    cnn = concatx ⊕concaty;
    O_{cnn} = Dense(64)(cnn);
    For i in range(2):
        Beign
        Parameters of Capsule Network:capsulenum=3,dimcapsule=16, routing=3
        End
    Input the features of Capsule network into self-Attention to obtain att
    $O_{capsule}$ = Dense(64)(att)
    A = $O_{cnn} \oplus O_{capsule}$
    A=1, Classification is correct,A=0,Classification is wrong
End

2) ANALYSIS OF THE EXPERIMENTAL RESULTS

The DHCN model proposed in this paper is compared with BiLSTM, CNN, Capsule, CNN-ComPooling, and Capsule-att on the basis of different text representation methods for the task of Chinese short text entity disambiguation. The experimental results are shown in Table.4.

For Word2vec (C + W + P)-BiLSTM and Word2vec (C + W)-BiLSTM, P is increased by 1.92%, R is increased by 3.96%, and F1 value is increased by 4.55% because the addition of position feature can help the model to perceive the position information of semantic vectors in the text and better understand the semantic relationships in the data. Compared with Word2vec (C + W + P)-BiLSTM, BERT-BiLSTM has a 9.44% higher P, 10.05% higher R, and 8.39% higher F1 because the word vector generated by BERT is dynamic, can model the polysemy phenomenon of a word, and can more accurately reflect the actual meaning of words in the current semantics. In the following experiments, BERT is used as the text vector representation.

Due to the short text length, when a model is trained, the semantic temporality has less effect. The BERT-CNN model uses different sizes of convolution kernels to extract important information in the sentence to better capture words. Compared with that of BERT-BiLSTM, the relationship between words increased by 1.96%, 1.04%, and 1.5% in terms of P, R, and F1 in BERT-CNN. BERT-CNN-ComPooling uses the commonly used fusion features of mean-pooling and max-pooling to represent the text. Max-pooling extracts the most significant features of the text. Mean-pooling considers all semantic information in the neighbourhood. The performance of the model is greatly improved compared with that of BERT-CNN. BERT-Capsule-att adds a self-attention mechanism to BERT-Capsule, which enables the model to more deeply understand the semantic information obtained by the Capsule model. Therefore, BERT-Capsule-att is 1.42%, 0.99%, and 1.15% higher than BERT-Capsule in terms of P, R, and F1, respectively.

The accuracy improvement of the entity disambiguation task of Chinese short text depends on how sufficient semantic information is obtained. To reduce the negative effects of semantic sparseness on short texts, the DHCN model proposed in this paper uses BERT as a text representation to capture semantic information through a two-layer capsule

TABLE 4. Comparison of experimental results.

Model	P/%	R/%	F1/%
Word2vec(C+W)Bi-LSTM	66.93	63.86	65.36
Word2ve(C+W+P)-Bi-LSTM	68.85	67.82	69.91
BERT-Bi-LSTM	78.29	78.32	78.30
BERT-CNN	80.25	79.36	79.80
BERT-CNN-ComPooling	84.72	84.12	84.29
BERT-Capsule	83.03	83.12	83.07
BERT-Capsule-att	84.45	84.11	84.22
BERT-DCHN	**88.14**	**87.98**	**88.04**

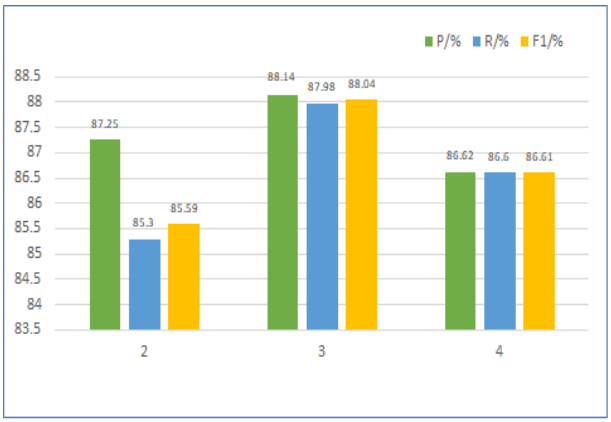

FIGURE 3. The number of capsules effect comparison.

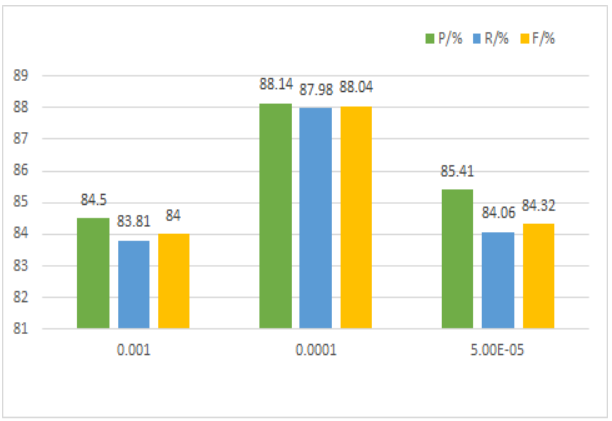

FIGURE 4. Comparison of the effects of different learning rates.

network, and adds a self-attention mechanism to further learn the feature information learned by the capsule network, instead of a simple flattening, which assigns weights to different semantics through self-attention to better obtain semantic dependencies. Furthermore, the CNN model, which is more suitable for short text NLP tasks, is applied to extract the text semantic information output by BERT with convolution kernels of different sizes (3, 4, 5), and the extracted semantic information is screened by merging max-pooling and the mean-pooling results. By merging the two pooling methods, the most significant features and all the semantic information in the neighbourhood can be effectively fused. Fusing the above semantic information in a dual-channel manner can effectively avoid the vanishing gradient problem of deep learning models with high depth. As seen from Table 4, the dual-channel hybrid network proposed in this paper has achieved better experimental results than the existing mainstream models in the task of Chinese short text entity disambiguation, which validates the effectiveness of the model.

3) ABLATION EXPERIMENT

a: THE INFLUENCE OF THE LEARNING RATE ON THE EXPERIMENTAL PERFORMANCE

To explore the influence of different learning rates on the performance of the model, learning rates of 0.001, 0.0001, and 0.00005 are used to conduct the experiments. The experimental results are shown in Figure.3.

As shown in Figure 4, when the learning rate is 0.0001, the DHCN model achieves the best effect, with P, R, and F1 values of 88.14%, 87.98%, and 88.04%, respectively. However, when the learning rate increases to 0.001, the results of the model decrease by 3.64%, 4.17%, and 4.04% for P, R, and F1, respectively, because when the learning rate is excessive, the model crosses the optimal value easily, which results in the model not converging for a long time and a poor training effect. However, when the learning rate is 0.00005, the results of the model decrease by 2.73%, 3.92%, and 3.72% in terms of P, R, and F1, respectively, because when the learning rate is too small, the model easily falls into local optima and cannot reach the global optimal point.

b: THE INFLUENCE OF THE NUMBER OF CAPSULES ON THE EXPERIMENTAL PERFORMANCE

To explore the influence of the number of capsules on the experimental results, this paper set the number of capsules to 2, 3, and 4. The experimental results are shown in Figure.4

As shown in Figure 5, the model performs best when 3 capsules are used. When the number of capsules is set to 2, the model achieves 87.25%, 85.3%, and 85.59% P, R, and F1, respectively, and the corresponding indicators decrease 0.89%, 2.68%, and 2.45%, respectively. This is because if the number of capsules is too small, the model will under-fit the data set. When the number of capsules is 4, the model achieves 86.62%, 86.6%, and 86.61% P, R, and F1, respectively, and the phase decreases by 1.52%, 1.38%, and 1.43% in the corresponding indexes. This is because too

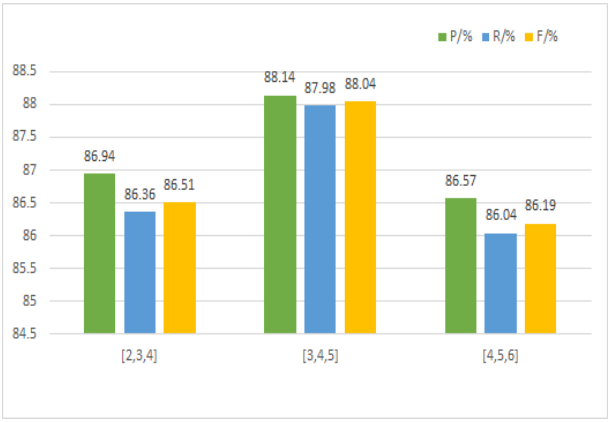

FIGURE 5. Comparison of the effects of different learning rates.

many capsules extract semantic information, which leads to over-fitting of the model and reduces the migration ability of the model.

c: THE INFLUENCE OF DIFFERENT CONVOLUTION KERNEL SIZES ON THE EXPERIMENTAL PERFORMANCE

To explore the influence of convolution kernel size on the performance of the model, this paper selects convolution kernels with sizes of [2, 3, 4], [3, 4, 5], and [4, 5, 6]. The experimental results are shown in Figure.5.

As shown in Figure 6, when the size of the convolution kernel is [3, 4, 5], the model achieves the best effect, with 88.14%, 87.98%, and 88.04% P, R, and F1, respectively. However, when the size of the convolution kernel is [2, 3, 4], the window is small, which makes the range of the receptive field smaller, and the receptive field contains less text information and limited features. Therefore, the model cannot fully express that features of the sentence, resulting in 1.2%, 1.62%, and 1.53% decreases in P, R, and F1 values. When the size of the convolution kernel is [4, 5, 6], because the change in the receptive field makes the global features obvious, considerable noise exists in the text semantic information obtained by the model, and the complexity of the model increases. The P, R, and F1 of the model decrease by 1.57%, 1.94%, and 1.85%, respectively.

V. CONCLUSION

Research on entity disambiguation technology is an important part of the construction of a knowledge graph. Accurate entity disambiguation results are the premise of building a knowledge graph. A high-precision knowledge graph can provide data support for NLP tasks such as intelligent question answering and information retrieval. In the task of Chinese short text entity disambiguation, due to the sparse semantics of Chinese short text, it is difficult for the model to learn sufficient semantic information, resulting in low accuracy. The dual-channel hybrid network proposed in this paper is a parallel model that combines a hybrid pooled CNN model with a capsule network model with a self-attention mechanism to fully learn semantic information. After merging the learned knowledge of the two models, the proposed approach uses unique identifiers to construct a text classification method to perform entity disambiguation tasks. The experimental results show that the DHCN proposed in this paper is superior to the existing mainstream models for Chinese short text entity disambiguation. Although our model has achieved state-of-the-art results, some problems remain to be addressed. For example, there are entities with the NIL identifier in the data. The next step is to judge the upper concept type based on the NIL type to solve the problem of entities of type NIL not being disambiguated, thereby improving the accuracy of disambiguation.

REFERENCES

[1] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity disambiguation for noisy social media posts," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Melbourne, VIC, Australia, 2018, pp. 2000–2008. [Online]. Available: https://www.aclweb.org/anthology/P18-1186

[2] P. Agarwal, J. Strötgen, L. del Corro, J. Hoffart, and G. Weikum, "DiaNED: Time-aware named entity disambiguation for diachronic corpora," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2. Melbourne, VIC, Australia, 2018, pp. 686–693. [Online]. Available: https://www.aclweb.org/anthology/P18-2109

[3] F. M. Suchanek, G. Kasneci, and G. Weikum, "Entity disambiguation method based on multifeature fusion graph model for entity linking," *Appl. Res. Comput.*, vol. 10, pp. 35–40, 2017.

[4] J. Wang, Z. Wang, D. Zhang, and J. Yan, "Combining knowledge with deep convolutional neural networks for short text classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2915–2921.

[5] G. Zhou, J. Zhao, K. Liu, "Open information extraction," *J. Chin. Inf. Process.*, vol. 25, no. 6, pp. 98–110, 2011.

[6] J. Moreno, R. Besancon, R. Beaumont, E. D'hondt, A.-L. Ligozat, S. Rosset, X. Tannier, and B. Grau, "Combining word and entity embeddings for entity linking," in *Proc. Eur. Semantic Web Conf.* Portoroz, Slovenia: Springer, Jan. 2017, pp. 337–352.

[7] S. S. Kataria, K. S. Kumar, R. R. Rastogi, P. Sen, and S. H. Sengamedu, "Entity disambiguation with hierarchical topic models," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 1037–1045.

[8] Z. Liu, Q. Lu, and J. Xu, "High performance clustering for Web person name disambiguation using topic capturing," in *Proc. 1st Int. Workshop Entity-Oriented Search (EOS)*, p. 1.

[9] A. Bagga and B. Baldwin, "Entity-based cross-document coreferencing using the vector space model," in *Proc. 17th Int. Conf. Comput. Linguistics*, vol. 1, 1998, pp. 1–7.

[10] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Disambiguating personal names on the Web using automatically extracted key phrases," *Frontiers Artif. Intell. Appl.*, vol. 141, pp. 553–557, Jan. 2006.

[11] Z. He, S. Liu, M. Li, M. Zhou, L. Zhang, and H. Wang, "Learning entity representation for entity disambiguation," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*. Sofia, Bulgaria, 2013, pp. 30–34.

[12] S. Hu, Z. Tan, W. Zeng, B. Ge, and W. Xiao, "Entity linking via symmetrical attention-based neural network and entity structural features," *Symmetry*, vol. 11, no. 4, p. 453, Apr. 2019.

[13] N. Gupta, S. Singh, and D. Roth, "Entity linking via joint encoding of types, descriptions, and context," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2681–2690.

[14] M. C. Phan, A. Sun, Y. Tay, J. Han, and C. Li, "Pair-linking for collective entity disambiguation: Two could be better than all," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 7, pp. 1383–1396, Jul. 2019.

[15] H. Huang, Y. Wang, C. Feng, Z. Liu, and Q. Zhou, "Leveraging conceptualization for short-text embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1282–1295, Jul. 2018.

[16] M. Yang, W. Zhao, J. Ye, Z. Lei, Z. Zhao, and S. Zhang, "Investigating capsule networks with dynamic routing for text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 3110–3119.

[17] M. Yang, W. Zhao, L. Chen, Q. Qu, Z. Zhao, and Y. Shen, "Investigating the transferring capability of capsule networks for text classification," *Neural Netw.*, vol. 118, pp. 247–261, Oct. 2019.
[18] W. Zhao, H. Peng, S. Eger, E. Cambria, and M. Yang, "Towards scalable and reliable capsule networks for challenging NLP applications," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 1549–1559.
[19] S. Broscheit, "Investigating entity knowledge in BERT with simple neural End-To-End entity linking," in *Proc. 23rd Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2019, pp. 677–685.
[20] A. M. Rinaldi, "A content-based approach for document representation and retrieval," in *Proc. 8th ACM Symp. Document Eng. (DocEng)*, 2008, pp. 106–109.
[21] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, "Feature selection via maximizing global information gain for text classification," *Knowl.-Based Syst.*, vol. 54, pp. 298–309, Dec. 2013.
[22] L. M. Abualigah and A. T. Khader, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering," *J. Supercomput.*, vol. 73, no. 11, pp. 4773–4795, Nov. 2017.
[23] D. Singh and B. Singh, "Hybridization of feature selection and feature weighting for high dimensional data," *Appl. Intell.*, vol. 49, pp. 1580–1596, Nov. 2018.
[24] T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[25] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162
[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*. [Online]. Available: https://arxiv.org/abs/1706.03762
[28] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, and X. Bai, "Named entity recognition using BERT BiLSTM CRF for chinese electronic health records," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–5.
[29] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," 2019, *arXiv:1904.05255*. [Online]. Available: http://arxiv.org/abs/1904.05255
[30] H. Xu, B. Liu, L. Shu, and P. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA, Jun. 2019, pp. 2324–2335. [Online]. Available: https://www.aclweb.org/anthology/N19-1242
[31] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751. [Online]. Available: https://www.aclweb.org/anthology/D14-1181
[32] Q. Xipeng, *Neural Networks and Deep Learning*. Beijing, China: China Machine Press, 2020.
[33] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
[34] S. Sabour, N. Frosst, and G. E Hinton, "Dynamic routing between capsules," 2017, *arXiv:1710.09829*. [Online]. Available: http://arxiv.org/abs/1710.09829
[35] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. Int. Conf. Artif. Neural Netw.*, 2011, pp. 44–51.
[36] N. Zhang, S. Deng, Z. Sun, X. Chen, W. Zhang, and H. Chen, "Attention-based capsule networks with dynamic routing for relation extraction," 2018, *arXiv:1812.11321*. [Online]. Available: http://arxiv.org/abs/1812.11321
[37] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, "Sentiment analysis by capsules," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 1165–1174.
[38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473
[39] Y. Kim, C. Denton, L. Lan, and A. Rush, "Structured attention networks," Feb. 2017, *arXiv:1702.00887*. [Online]. Available: https://arxiv.org/abs/1702.00887
[40] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.

LITING JIANG received the bachelor's degree from the Lanzhou University of Finance and Economics, in 2017. She is currently pursuing the M.S. degree with the College of Information Science and Engineering, Xinjiang University, China. Her research interests include natural language processing and knowledge graph.

GULILA ALTENBEK received the M.S. degree in computer application from the University of Science and Technology Beijing, in 1996, and the Ph.D. degree in computer application technology from the Harbin Institute of Technology, China, in 2017. Since 1985, she has been studying in natural language processing. She is currently a Professor with Xinjiang University. Her research interests include natural language processing, artificial intelligence, and knowledge graph.

DI WU received the bachelor's degree from the Hubei University of Medicine, in 2018. He is currently pursuing the M.S. degree with the College of Software Engineering, Xinjiang University, China. His research interests include natural language processing and knowledge graph.

YAJING MA received the bachelor's degree in engineering from Zaozhuang University, in 2017, and the master's degree in engineering from Northwest Minzu University, in 2019. She is currently pursuing the Ph.D. degree with the College of Information Science and Engineering, Xinjiang University, China. Her research interests include natural language processing and knowledge graph.

HAYINAER AIERZHATI is currently pursuing the Bachelor of Engineering degree with the School of Computer Science and Technology, University of Science and Technology of China, China. His current research interest includes computer application.