

THEORY

Multi-Level Attention Based Coreference Resolution With Gated Recurrent Unit and Convolutional Neural Networks

BIANBADROMA^{1,2}, NGODRUP¹, ERPING ZHAO¹, YUHAO WANG¹,
AND YAKUN ZHANG¹

¹College of Information Engineering, Xizang Minzu University, Xianyang, Shaanxi 712082, China

²Tibet Net-Cloud Sci-Tech Company Ltd., Tibet, Lhasa 850001, China

Corresponding author: Erping Zhao (xdzep@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61762082, and in part by the Science and Technology Project of Tibet Autonomous Region under Grant XZ202001ZY0055G.

ABSTRACT Aiming at the diversity of the entity mentions in the field of animal husbandry in Tibet, the reference resolution model Att-GRU-CNN based on multi-level attention mechanism is proposed. The model uses GRU network for global semantic feature learning and knowledge memory, and uses CNN network to further extract local high-level semantic features. A word-level attention layer is added on the GRU hidden layer, the feature vector of entity mention does dot product operation with the feature vector of each word, and the result is normalized and used as the weight value distribution of the words, thus, the prior knowledge that the mention is the most important in the context is given to the network; A sentence-level attention layer is added on the CNN convolution layer, the convolutional layer output and entity name library matrix do association operation, and the result is used as the weight value distribution of the sentence, so as to strengthen the relation between the sentence where the mention is located and the entity scientific name. At the same time, the data enhancement technology is used to improve the generalization ability of the model. Finally, the ablation experiment on Tibetan animal husbandry dataset verified the effectiveness of each component of the model and the superposition effect after combination; the performance comparison experiment is carried out on public datasets MUC, B3 and CEAF_{φ4}. The experimental results show that this model has a significant improvement over other models in the accuracy rate, recall rate and F1 score.


INDEX TERMS Mention diversity, coreference resolution, attention mechanism, GRU-CNN, data enhancement.

I. INTRODUCTION

The entity mention diversity means that a named entity can be expressed in many ways, that is, the name of an entity has multiple different mentions. Coreference resolution aims at determining whether two or more mentions in a text refer to the same real-world entity, it normally needs to complete two subtasks; one subtask is to identify the potential entity mentions which appear in a given context, and the other is to group mentions into different clusters such that mentions in each cluster point to the same target entity. Coreference

resolution is an import task in Natural Language Processing(NLP), it is used in various natural language processing pipelines, such as machine translation, question answering and text simplification [1].

With the explosive growth of network information, information in various fields is diversified, coupled with the flexibility of natural language and the diversity of expressions. Chinese has both the standard language and the dialect, as well as the difference between the written language and the spoken language, which leads to the widespread phenomenon of multiple mentions in the Chinese entity. For example, the entity "Artemisia annua" in the Tibet animal husbandry dataset has many mentions such as

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés .

”mugwort”, ”vetch grass” and ”twelve younger sisters”, etc. Another example, people’s names have both scientific names, nicknames, pseudonyms, and previous names and so on. This phenomenon that an entity has multiple mentions leads to a lot of manpower to resolve the multiple mentions of entities before relationship extraction and knowledge graph construction. The study of coreference resolution is of great significance, and in the process of knowledge graph construction and knowledge reasoning, multiple mentions of entity cause the same object in the real world to appear repeatedly, resulting in confusion in entity links and errors in knowledge reasoning. Coreference resolution has become one of the most popular research fields for detecting the same entities in various practical scenes [2]. The earlier research on coreference resolution is based on rules and statistics [3], [4]. However, whether rule-based or statistical-based, they only aim at the limited content of the same text, and the accuracy will decrease significantly with the increase of the amount of data, and these approaches require a large amount of human effort and time to create rules to solve the coreference resolution task. The coreference resolution model based on machine learning uses context vectors to constituting a new form of language representation, in addition, a combination of dependency grammar, sentiment ontology and so on [5]. Yuan et al. [6] proposed a coreference resolution model based on active learning, and the modified model solves the problem that coreference resolution models have trained on one dataset may not be transferred to new low-resource domains. The solution to this problem is to sample a small subset of data for annotators to label, but, labeling more spans across different documents reduces model performance. The shortcomings of the machine-learning approaches are that the features of the training data are often local, without considering the global dependencies and semantic features, and may violate the transitivity of the co-referential equivalence relation. In recent years, the research mainly focuses on reference resolution based on deep learning. For example, an actor-critic-based neural coreference resolution system can better achieve both mention detection and mention clustering [7], its contribution is to utilize the BERT model obtain better span representations, thus improve the reference resolution model performance. The BERT-SRU-based Pointer Networks for coreference resolution model was proposed that leverages the linguistic property of head-final languages, and combines with Knowledge Distillation [8]. This model is only suitable for strongly head-final languages and head-initial languages, but it is not very effective for complex Chinese language.

The focus of our research is to determine whether an entity in a context appears two or more mentions and then make them refer to unique standard mention of the entity (scientific name). For example, how ”naked barley” and ”yuan wheat” in context refer to ”highland barley” (scientific name) is our work. A deep learning combined with word embedding, neural network, attention mechanism and other technologies has a strong ability to learn semantic information, and

improves the overall performance and generalization ability of the model, especially prominent in the tasks of co-referential resolution, so we propose an entity reference resolution model Att-GRU-CNN based on multi-level attention mechanism. We use the language preprocessing model Bert (bidirectional encoder representation from transformers) to do words embedding representation of the text, the GRU (Gate Recurrent Unit) is responsible for knowledge memory and text feature learning, and word-level attention layer is added to GRU to emphasize the importance of entity mention. The Convolutional Neural Networks (CNN) extract local high-level semantic features, at the same time add sentence-level attention layer to CNN, so as to strengthen the relation between the sentence in which the entity mention is located and corresponding the entity scientific name in the scientific name library.

The main contributions of this paper are as follows: (1) we propose a novel word-level attention mechanism, the feature vector of entity mention does dot product operation with the feature vector of each word, and the result is normalized and used as the weight value distribution of the words, thus, the prior knowledge that the mention is the most important in the context is given to the network, so as to enhance the model’s attention to the mention. (2) We propose a novel sentence-level attention mechanism, the output of the CNN convolution layer is multiplied by the entity scientific name library matrix to obtain the relationship matrix between the sentences and the entity scientific name library, this relationship matrix is input into the scoring function to calculate each sentence weight value, so as to enhance the relationship between the sentence where the entity mentions and the corresponding entity scientific name (the standard mention) in the scientific name library. (3) We combine GRU with CNN as a training model, which not only solves the problem of long-distance dependence, but also extracts local high-level semantic features, and further improves the semantic feature learning ability of the model.

The remainder of this article is organized as follows. The next section analyzes the related work on the reference resolution models and methods. Section III presents the implementation of our models and methods. Section IV illustrates and comparative analyses the experiment results. Finally, the conclusion and prospect of this article is drawn in section V.

II. RELATED WORK

With the development of natural language processing research, scholars have made a series of researches on coreference resolution, since early rule-based to statistical ones until to machine and deep learning-based [9]. A rule approach depends on some hand-crafted rules based on syntactic, semantic features and a more robust evaluation [10]. Because the rule approach requires a lot of manpower and material resources to craft the rule template, this approach is gradually eliminated. The machine-learning approaches regard reference resolution as a binary classification and dual or

dualistic classification and learning to rank problem, and then reference resolution tasks are completed by clustering algorithm, statistical algorithm, EM algorithm, LDA algorithm and other algorithms. Lij et al. [11] proposed active learning for coreference resolution using discrete annotation, this model improves upon pairwise annotation for active learning, and combine with a novel mention clustering algorithm to improve machine learning performance in in coreference resolution. A method of cross-document coreference resolution was presented based on unsupervised learning [12], which resolves semantically complex mentions, more loose coreference relations and mentions in the “wild” of political news articles, it includes resolution of named entity, resolution of groups of persons and resolution of event entity phrase.

Compared with the machine learning methods, the deep learning neural networks can realize layer-by-layer abstraction by using distributed representation, which is good at learning and extracting high-level semantic features of specified task. Guarasci et al. [10] and others proposed an Italian coreference resolution system based on the end-to-end architecture of neural networks, with ELECTRA as the language pre-training model. The neural network models resolve Russian coreference resolution with word vector representation and using semantic information in a text [13], [14]. In the thirteenth reference, the coreference resolution model based on feedforward neural network with word embedding representation consists of two main modules: the mention-pair encoder and the mention-ranking layer. In the fourteenth reference, the coreference resolution model is based on a Bidirectional long short-term memory layer neural network, and integrates features derived from open-source semantic information. These two models have achieved good results in the task of Russian reference resolution, and proved the importance of language models and semantic features. Kopyt et al. [15] proposed a fully connected neural network method to complete the coreference resolution of Persian, and mainly contributed to extract and fuse some handcrafted features, word embedding features and semantic features, F-score is 64.54%.

In order to solve ambiguous pronouns in the field of coreference resolution, the pre-trained BERT and PyTorch helper bot is used to embed context, and then applies a custom made Multilayer Perceptron as a classifier to predict the probability in each class to identify the correct mention of the target pronoun [16]. In recent years, coreference resolution has received a sensibly performance boost exploiting pre-trained language model BERT generating different span embedding representations as a model input [7], [8], [10], [16], [17]. Attention mechanism was first used in the field of machine translation, and now it has been widely used in the task of reference resolution, it can make the neural network pay attention to the key words and sentences selectively, and enhance the semantic representation of these words and sentences, and so as to improve the network learning effect. The introduction of it aims to strengthen the representation of key words and sentences relative to other elements.

The end-to-end coreference resolution system was improved by using a biaffine attention model to get antecedent scores for each possible mention [18]. Cheng et al. [19] introduced a decomposable attention neural network model DANGL with global inference mechanism based on remote and local information to document-level event coreference resolution. A neural coreference resolution model employed mutual attention to take into account the dependencies between spans and their proceeding spans directly, and used attention mechanism to capture global information between spans and their proceeding spans [20]. Jiang et al. [21] use a Graph Attention Network to incorporate syntactic and semantic structures of sentences, which allows the model to selectively incorporate information from its neighbour nodes, so as to improve the model co-reference resolution effect.

III. MODEL

GRU network is good at dealing with timing sequences and extracting the contextual semantic features of non-contiguous words in statements, that is, long-distance semantic features capture, but GRU is not good at capturing key local features. CNN can extract the context semantic features of words within the convolution kernel through the convolution layer, and CNN is good at extracting the local semantic features between continuous words. Therefore, we combine GRU and CNN to build a training model, giving full play to the advantages of the two networks and complementing their shortcomings, so as to improve the semantic feature extraction ability of Att-GRU-CNN. BERT is a language pre-trained model that can transfer learning. It can capture the deep and bidirectional information between words in a sentence, and express the vector of the whole sentence with the vector of [CLS] token in the output layer, so this paper uses BERT for embedding representation. As shown in Fig 1, the input sequence of BERT is $W = \{w_1, w_2, \dots, w_n\}$, and the sequence represented by BERT embedding is $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, which will be used as the input of GRU network. GRU is used for learning and memory the features of the input sequence, the output sequence of hidden layer after GRU training is $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$, and introduce the word-level attention mechanism for the hidden layer. The output of GRU hidden layer which was multiplied by word-level attention weight coefficient is taken as the input of CNN, the convolution layer of CNN is responsible for extracting local semantic features, and a sentence-level attention mechanism is introduced between the convolution layer and the pooling layer, so that sentences containing different entity mentions can be associated with Chinese scientific names in the scientific name library. Finally, feature fusion is carried out in maximum pooling, and the overall structure of the model is shown in Figure 1.

A. GRU NETWORK

GRU and long-short term memory (LSTM) are all belong to recurrent neural networks (RNN). The difference is that GRU combines the forgetting gate and the input gate into an update

gate, so GRU has fewer training parameters, easier converges and faster training speed. At the same time, GRU as a variant of LSTM can also deal with the long-distance dependence problem of RNN, accurately predict through the retained memory information, and train semantic information that conforms to the corpus of a specific domain. GRU training text has strong timeliness and relevance of non-continuous contextual information. Therefore, we use GRU for feature learning and memory. The internal structure of GRU is shown in Figure 2.

Hadamard product operator. z_t and r_t represent the status of update and reset gates. h_t represents the state of the hidden layer at time t . t is the state of the candidate hidden layer at time t . x_t represents model input. GRU determines its output through two moments before and after, so that it can well learn the semantic features of mention context when processing sequential text. But it cannot pay special attention to the mention, so we add word-level attention mechanism on the basis of GRU to enhance the representation of mentions relative to other words.

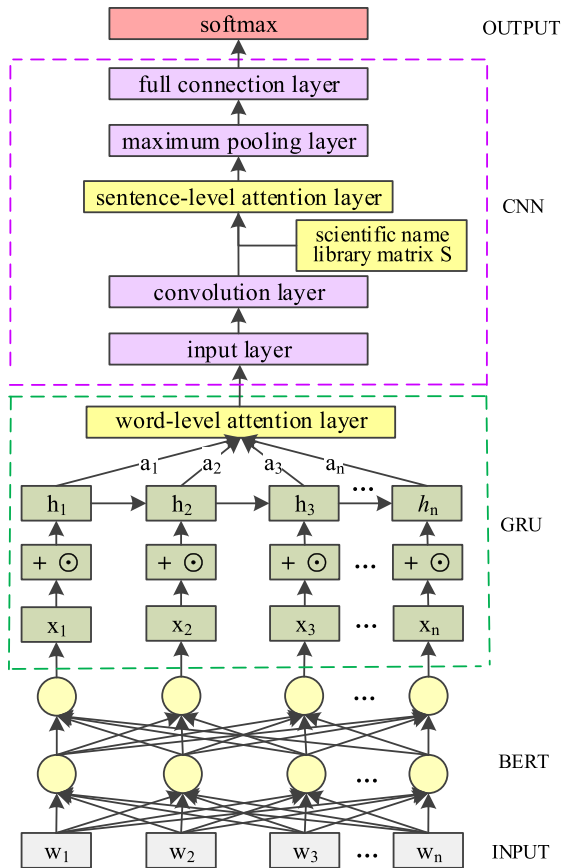


FIGURE 1. Architecture of our proposed model Att-GRU-CNN for the coreference resolution task.

In Figure 2. z_t and r_t represent update and reset gates respectively. Because of this special gate structure, GRU can precisely choose the information to transmit to achieve the purpose of controlling the information. The model formulas are defined as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (1)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (2)$$

$$\hat{h}_t = \tanh(W_h x_t + U_h (r_t \odot h_{t-1})) \quad (3)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \quad (4)$$

where W_z , W_r , U_z , U_r represent the weight matrix of the update and reset gates. W_h , U_h represents the weight matrix of the candidate hidden layer. $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent function, \odot represents

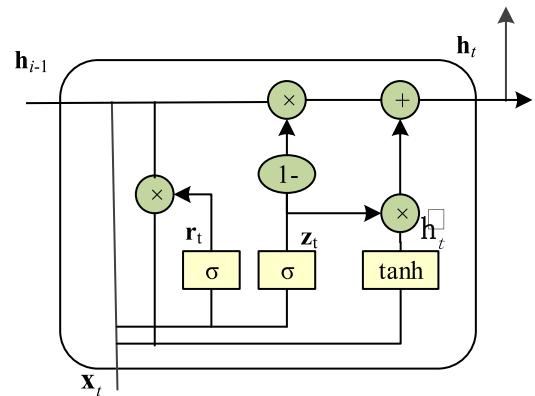


FIGURE 2. GRU internal structure. z_t and r_t respectively represent the update gate and reset gate in the figure. Combining the input x_t with the hidden state h_{t-1} passed from the previous node, the GRU takes the output of the current hidden node and the hidden state is passed to the next node.

B. WORD-LEVEL ATTENTION MECHANISM

The introduction of word-level attention mechanism in this paper aims to highlight the importance of entity mentions relative to other words in sentences, enhance the information expression of entity mentions, and make the network model can pay key attention to entity mentions in the training process. Each mention of an entity is quite different in the representation of context, for example, other mentions of the entity "cordyceps sinensis" include "summer grass winter insect" and "insect grass" and "winter grass insect" and so on, furthermore the neural network's attention to important words will decrease with the complexity of data, in order to make up for this deficiency, scholars usually add attention layer to the neural network, and most of them are self-attention layer. The disadvantage of self-attention layer is that the neural network pays equal attention to each word in the sequence at the beginning of training, only in the iterative training process, the neural network gradually pays different attention to different words. In this paper, the purpose of adding word-level attention layer is to endow GRU network with prior knowledge of different importance of each word in a sentence, and prior knowledge that the mention is the most important in a sentence, so that the neural network can very focus on entity mentions at the beginning of training, thereby enhancing the important role of the mention features in the training process. Wang et al. [22] and others integrated entity

information and its location information into the weight value distribution function to calculate the weight value of each word. Guo et al. [23] and others multiplied different words to modify the context representation weight of the original word, increased the weight value of the words that exist in the path, and achieved better results.

In this paper, the principle of word-level attention mechanism and the calculation process of weight value are as follows. The embedded expression sequence of the sentence is $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, and the output sequence of the hidden layer calculated by formula (4) is $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$. In order to highlight the importance of entity mentions in the sequence of the hidden layer, we take the Eigenvector \mathbf{h}_e of the entity mention in the hidden layer as the benchmark, do the dot product operations one by one with the eigenvector of other words, and normalize the dot product values, this result is the attention distribution of each word, so that the entity mentions can obtain the largest initial weight coefficient. We use the position information of the dataset tagging to locate the Eigenvector of entity mention in \mathbf{H} sequence, and the Eigenvector of entity reference is recorded as \mathbf{h}_e , and then use \mathbf{h}_e and other word eigenvector of the hidden layer to do the dot product operation one by one, and the dot product value is normalized as the weight value of each word. Its calculation is defined as:

$$a_i = \frac{\exp(\mathbf{h}_e^T \mathbf{h}_i)}{\sum_{i=1}^n \exp(\mathbf{h}_e^T \mathbf{h}_i)} \quad (5)$$

where a_i represents the initial weight value given to the i -th column vector of the hidden layer $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$. Since the weight value corresponding to the entity mention is the largest, GRU can pay key attention the entity mention, the weight distribution is defined as:

$$\mathbf{H}^* = \mathbf{H} \otimes \mathbf{a}^T \quad (6)$$

where \mathbf{a}^T represents the transposition of the weight value vector, \otimes and represents element-by-element multiplication, and finally, the matrix \mathbf{H}^* represents the hidden layer output with attention.

In the past, the self-attention mechanism paid equal attention to each word in the sequence at the beginning of model training, and our word-level attention mechanism can give GRU network the prior knowledge that entity mention is very important in sentence in advance, that is, each word in the sequence is given a different weight value before training, and the mention is given the largest weight value. this enables GRU network to focus on entity mention at the beginning of training, so as to better extract semantic features of mentions in context. Finally, the output of GRU hidden layer with word level attention is input to CNN network, so that CNN network can further extract high-level semantic features.

C. CNN NETWORK

The unique gating mechanism of GRU will filter information, so that some key local features can't be obtained. Moreover,

the reference resolution task in this paper focuses on the different mentions of entities in sentences. so our work requires that the model not only has a certain long-distance memory ability, but also can extract the potential local features of mentions. CNN is a feedforward neural network with strong local feature extraction ability, local semantic features of reference in sentences using CNN can be extracted well. We combine the respective advantages of GRU network and CNN network to construct the GRU-CNN model, which makes the model good at both long-distance feature capture and local feature extraction. We use the GRU hidden layer output $\mathbf{H}^* \in \mathbb{R}^{i \times j}$ as the CNN network input, where i represents the number of word vectors in the sentence and j represents the word vector dimension. The convolution kernel is expressed as $\mathbf{V} \in \mathbb{R}^{k \times j}$, where k represents the convolution kernel size, and k value is set as 3, 4 and 5, and the number of filters is set as 128. The output of convolution layer is as follows:

$$\hat{\mathbf{H}}_m^* = f(\mathbf{V} \odot \mathbf{H}_{m|m+k-1}^* + \mathbf{b}) \quad (7)$$

where \odot represents the convolution operator, \mathbf{b} represents the bias offset, f represents is a nonlinear activation function, here we utilize the leaky ReLU activation function, $\mathbf{H}_{m|n}^*$ represents the matrix of eigenvectors from m to n , $\hat{\mathbf{H}}_m^*$ represents the eigenvalue of the m -th word, and we further extract the local semantic features in the sequence through convolutional layer.

D. SENTENCE-LEVEL ATTENTION MECHANISM

The sentence-level attention mechanism we introduce aims to establish the association of sentences containing different mentions of the entity with the entity scientific name library, thereby enhancing the semantic connection between these sentences and their corresponding entity scientific name. Wei et al. [24] and others constructed a two-layer word-level and sentence-level attention mechanism and emphasized the synergistic effect of the two, which achieved good results. Since it is not enough to rely on a single word to express the meaning of the entire sentence. We introduce the sentence-level attention mechanism to strengthen the relationship between sentences containing an entity mention and this entity scientific name in the scientific name library, while weakening the association between these sentences and unrelated entity scientific names.

The sentence-level attention layer is added between the CNN convolution layer and the maximum pooling layer, which establishes the correlation between the convolution layer output $\hat{\mathbf{H}}^*$ and the entity scientific name library matrix \mathbf{S} , in the learning process, the reinforcement model specially focuses on the sentences where the entity reference is located. We input all the scientific names in the entity scientific name library into the pre-training model BERT in sequence, and generate the word embedding representations of the scientific name sequences, thereby further obtain the scientific name library matrix \mathbf{S} . The relation matrix \mathbf{G} between the sentence and the entity name library is obtained

by multiplying the output $\hat{\mathbf{H}}^*$ of CNN convolution layer with entity name library matrix \mathbf{S} , and then the relation matrix \mathbf{G} is inputted into the softmax function to further calculate the weight value of each sentence, the weight value is calculated as follows:

$$\mathbf{G} = \hat{\mathbf{H}}^{*T} \mathbf{U} \mathbf{S} \quad (8)$$

$$\beta_j = \frac{\exp(\mathbf{G}_j)}{\sum_{j=1}^m \exp(\mathbf{G}_j)} \quad (9)$$

where $\hat{\mathbf{H}}^*$ represents the convolutional output of the CNN, \mathbf{U} is the auxiliary matrix we introduce, and \mathbf{G}_j represents the relationship matrix between the j -th sentence and the entity scientific name library, β_j represents the weight value of the j -th sentence.

Then the sentence-level attention weight distribution is automatically assigned to the output $\hat{\mathbf{H}}^*$ of the CNN convolutional layer, which is distributed as follows:

$$\mathbf{H}^{**} = \hat{\mathbf{H}}^* \otimes \beta^T \quad (10)$$

where β^T represents the transpose of the sentence-level weight vector. The convolutional layer $\hat{\mathbf{H}}^*$ multiply by the transpose β^T of the sentence-level weight vector, and the convolution layer output \mathbf{H}^{**} with weight value is obtained. The pool layer removes redundant information from \mathbf{H}^{**} , extracts and compresses the local semantic features of the entity mention. The feature vector \mathbf{H}^{**} processed by the pooling layer carry the local high-level semantic features of the entity mention. The pooling layer adopts the maximum pooling method, which is defined as follows:

$$\mathbf{H}^{**} = \hat{\mathbf{H}}^* \otimes \beta^T \quad (11)$$

E. THE MODEL TRAINING

The essence of the reference resolution is a binary classification problem; the entity scientific name is used as a classification label to judge whether the entity mention in the sentence matches its scientific name in the scientific name library. We choose the Softmax function as the classifier, which is defined as:

$$p = s(\mathbf{M}\hat{\mathbf{H}}_i^{**} + \mathbf{b}_i) \quad (12)$$

where $\hat{\mathbf{H}}_i^{**}$ is calculated by the formula (11), \mathbf{b}_i represents the bias vector, s represents the softmax function, \mathbf{M} represents the weight matrix of the output layer, p is the matching probability of output, and it is used to judge which entity scientific name the current mention belongs to. The loss function adopts the minimum cross-entropy loss function, which is distributed as follows:

$$F = -\frac{1}{|d|} \sum_{i=1}^{|d|} [p_i \cdot \log(1 - p_i) \cdot \log(1 - p)] \quad (13)$$

where $|d|$ represents the number of the samples in the dataset, and p represents the matching probability of output. p_i represents the probability that this mention matches the current entity scientific name successfully.

IV. EXPERIMENTAL RESULTS

A. DATASET

The experimental data includes public dataset and self-built dataset, among which the data of the model performance comparison experiments use MUC, B3 and CEAF Φ_4 public datasets. The dataset of the Tibet Animal Husbandry is used to do the ablation experiment. The dataset is an achievement of the National Natural Science Foundation of China (61762082), there are 20163 sentences in this dataset, and these sentences are divided into training dataset, validation dataset and testing dataset with a ratio of 7:2:1. The entity scientific name library is composed of entity scientific names described from Baidu Encyclopedia pages by entity link technology. The construction process of the entity scientific name library is as follows: first, the training dataset is labeled with entities, and the labeled entities are regarded as entities to be matched; secondly, the entity to be matched is used for entity link to obtain the information on the encyclopedia page, and the knowledge description information in Baidu encyclopedia is saved, and the knowledge description information is filtered; third, we take each descriptive information item named "Chinese standard name" or "scientific name" as the corresponding scientific name of the entity linked to the entity; finally, all entity scientific names are saved in the library.

B. ENHANCEMENT SELF-BUILT DATASET

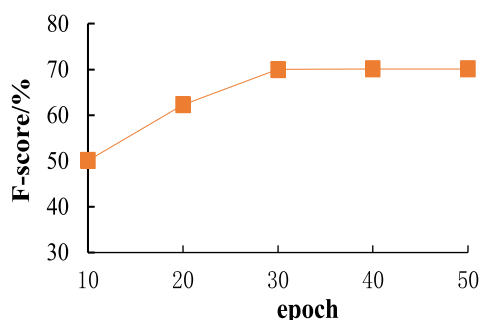
As the entity names in the field of animal husbandry in Tibet are influenced by Tibetan translation or transliteration, which leads to many entity names in this field to be diversified, that is, an entity have multiple mentions. For example, the animal with the Chinese scientific name of "yak" also has other names, such as "woolly rhinoceros", "Tibetan cattle", "horsetail cattle", "pig sound cattle", which are the mentions of this entity. In order to make the training data more sufficient to improve the generalization ability of the model, we use data enhancement technology to preprocess the experimental data so that the Experimental datasets cover all the mentions of every entity through the enhancement technique. We cross-replace the mentions in the sentences to produce new sentences and supplement them to the self-built dataset. The self-built dataset after data enhancement is 57,260 sentences, and the example of data enhancement is shown in Table 1.

C. OPTIMAL PARAMETERS

We use BERT pre-trained models for word embedding, hidden layer hidden_size is set as 768, and maximum length of text seq_len is set as 512, we use Adam optimizer [25]. Regularization method Dropout is used to reduce overfitting, and the learning rate is set to 0.01, the dropout rates of GRU and CNN are set to 0.5. In the training process, the best number of training rounds is obtained by K-fold cross validation method, as shown in Figure 3, the model tends to converge at 40.

TABLE 1. Table shows the example of data enhancement and the rule of cross-replacing the mention in sentences.

A yak is mainly distributed in Tibet and Qinghai.
A horsetail cattle is mainly distributed in Tibet and Qinghai.
A Tibetan cattle is mainly distributed in Tibet and Qinghai.
A pig sound cattle is mainly distributed in Tibet and Qinghai.
Why is a yak important to the Tibetan people?
Why is a horsetail cattle important to the Tibetan people?
Why is a Tibetan cattle important to the Tibetan people?
Why is a pig sound cattle important to the Tibetan people?

**FIGURE 3.** Figure shows that the optimal number of training rounds is 40 in our experiment, it is obtained by K-fold cross validation method, and the K value is 5.

The reason for using the K-fold cross validation method is to prevent the model from over fitting and improve the performance of the model on new data. The principle is to randomly divide the original data into K parts, with one part reserved for the test set and the rest ($K-1$) for the training set. After K times of cross validation, the average value is taken as the model evaluation standard, and the K value in this paper is 5. We uses Precision, Recall, and F1 (F-score) as performance indicators.

D. ABLATION EXPERIMENT

In order to verify the effective role of the single neural network model, the combined model of two neural networks and the multi-level attention mechanism respectively in the experiment, and secondly, to compare the effects of our attention mechanism and self-attention on the performance of the model, we specially conducted the ablation experiment. The specific settings of the experiment are as follows: 1) A single GRU model; 2) GRU-CNN model without attention mechanism; 3) SAtt-GRU-CNN model with self-attention mechanism; 4) Att-GRU-CNN model with our two-level attention mechanism. In order to reflect the advantages of the model used in this paper, the training data of the above model all adopts the data-enhanced data, and the BERT pre-trained model is used for embedded representation of the data. The ablation experimental results are shown in Table 2.

Table 2 shows that the two-level attention mechanism model we propose performs best. The effect of the GRU-CNN

TABLE 2. Ablation experimental verify the effective role of the single neural network model, the combined model of two neural networks, the word-level and sentence-level attention mechanism respectively in the experiment.

Model	P	R	F1
GRU	56.8	59.1	60.4
GRU-CNN	62.9	65.6	64.7
SAtt-GRU-CNN	66.7	68.7	68.2
Att-GRU-CNN	70.4	72.3	72.8

model that introduces the attention mechanism is significantly better than the GRU-CNN model without the attention mechanism. In the case that the model does not introduce the attention mechanism, the entity mention cannot highlight its importance relative to other words, the model can't learn the characteristics of the relationship between the mention sentence and its corresponding scientific name, the model can only learn long-distance information through GRU and extract local features through CNN, which leads to the low performance of the model. At the same time, the performance of a single GRU model is much lower than that of GRU-CNN model which combines GRU network with CNN network. Because a single GRU model lacks the extraction of local high-level semantic features, which leads to incomplete semantic feature extraction, the performance of a single model is the worst. It is that two models with attention mechanism were compared of SAtt-GRU-CNN and Att-GRU-CNN, due to the introduction of self-attention mechanism, the initial weight value assigned to each word in the training sequence assigned by the model is equal, which is constantly modified in the iterative training process. Our word-level attention mechanism gives the model the prior knowledge that the entity mention is very important in the sentence in advance, so that the initial weight of each word is unequal, and the initial weight value of mention is the maximum, thus strengthening the model's attention to the mention. Our sentence-level attention mechanism enhances the model to extract the information about the relationship between mention sentences and their corresponding entity scientific name, so that CNN network can better capture the features of different mention phrases in sentences, thus improving the model's ability to resolve entity mention and improving the matching degree between different mentions and their corresponding scientific names.

E. MODEL PERFORMANCE COMPARISON

We run experiments on the datasets of MUC, B^3 , CEAF $_{\phi_4}$, our model Att-GRU-CNN was respectively compared with the models without attention layer such as BiLSTM [26], CorefDPR [27], c2f-coref+ELMo [28], CorefQA+SpanBERT+base [29], and the models with attention layer such as JONA [20] and core-HGAT+SpanBERT [21]. In order to comprehensively evaluate models, we report the Precision,

TABLE 3. Model performance comparison. The table shows scores for different case studies.

Model	MUC			B^3			CEAF _{ϕ_4}			
	P	R	F1	P	R	F1	P	R	F1	Avg.F1
BiLSTM	77.70	63.91	70.13	70.00	54.01	60.98	40.34	57.74	57.74	62.95
CorefDPR	75.1	71.9	73.2	68.1	61.7	66.2	62.9	60.4	62.1	66.4
c2f-coref+ELMo	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
CorefQA+SpanBERT+base	85.2	87.4	86.3	78.7	76.5	77.6	76.0	75.6	75.8	79.9
JONA	82.3	79.5	80.9	73.6	69.4	71.4	69.6	66.8	68.2	73.5
coref-HGAT+SpanBERT	86.8	86.5	86.3	80.0	79.7	79.8	78	75.9	76.9	81.1
Att-GRU-CNN(ours)	85.3	86.8	86.6	78.9	79.6	80.3	76.7	76.2	77.3	81.4

Bold font indicates ours result. hidden layer hidden_size is set as 768, and maximum length of text seq_len is set as 512. The learning rate is set to 0.01, the dropout rates of GRU and CNN are set to 0.5, and the optimal number of training rounds is 40 in our experiment.

Recall and F1 score of the three metrics respectively and apply the average F1 score of them to be main evaluation, the experimental results are shown in Table 3.

The results in Table 3. show the three metrics of performance of our model and several successful models mentioned in the literature, and our model performs the best in these three metrics. Table 3. is divided into three parts: in the first one lists the four evaluation metrics of several models without attention layer, in the second one lists the metrics of models with attention layer, and in the third one is the metrics of our model with two-level attention layers. As a whole, the performance of models with attention mechanism far exceeds that of models without attention mechanism. It further proves that attention mechanism plays an important role in the task of coreference resolution based on neural network model, which is also fully verified in our ablation experiment.

The attention mechanism of model JONA is based on the coreference links between i and candidate j in context, and then this attention is given to the model. This attention mechanism, like the word-level attention mechanism in this paper, belongs to giving prior knowledge to the model in advance, so that the model will pay more attention to the words in a context at the beginning of training, and focusing on words of important position or words that play a key role in the task of the coreference resolution. Because the model JONA uses the position of candidate words in the context, but the candidate words themselves are uncertain, that is, the coreference links between i and candidate j in context are often not fixed, this leads to the uncertainty of the attention knowledge artificially given to the model, so the performance of the model JONA is slightly lower. The attention mechanism of the model core-HGAT+SpanBERT uses the graph attention network to propagate syntactic and semantic information to basic token nodes. For a node i , the attention mechanism allows it to selectively incorporate information from its neighbour nodes, so this attention mechanism belongs to self-attention. Since the self-attention mechanism is a rule gradually formed in the process of model training, the model pays equal attention to

each word at the beginning of training, so the performance of the model is also slightly lower.

F. ANALYSIS OF SELF-BUILT AND PUBLIC DATASETS

The results in Table 2. and Table 3. show that the performance of our model on public datasets is much higher than that of self-built dataset. One reason is that some entities in the self-built dataset have multiple mentions. For example, the mentions of the entity "blue sheep" includes "bharal", "stone sheep", "argali", "owe that", "Nawa", "Gongna" and so on, the entity and the mentions are one-to-many mapping, while CEAF _{ϕ_4} is based on entities, the entity and the mention is one-to-one mapping. The second reason is that MUC and B^3 are based on links and computed by measuring the common coreference links between gold-standard mentions/entities and mentions/entities that refer to them, the self-built dataset is based on the semantic features of the mention in the context, and the relationship between it and the corresponding scientific name in the entity scientific name library. The third reason, the entity types of self-built dataset are far more than those of these public datasets. They are not only people's names, place names, organizations, but also forage grass, grassland, animals, diseases, pests and diseases, and so on, more entity types will inevitably lead to that the task of coreference resolution becomes complicated and difficult. Due to the above reasons, the performance of our model on self-built dataset is lower than that on these public datasets.

V. CONCLUSION

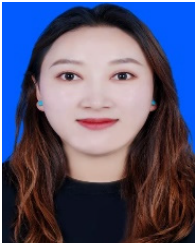
Aiming at the problem of multiple mentions of some entities in the field of animal husbandry in Tibet, this paper proposes a coreference resolution model Att-GRU-CNN, which combines multi-level attention mechanism and multiple neural networks, in order to make multiple mentions of any entity refer to its unique scientific name. We propose a new method for weight calculation of word-level and sentence-level attention mechanisms, which has the advantage of giving each word in the sequence a different initial weight value size, so

that the neural network model pay different attention to each word in the sequence at the beginning of training, especially focusing on different mentions of entities to better capture important semantic features. In this paper, the sentence-level attention mechanism is added to the CNN network, strengthens the relationship between sentences in which the different mentions of the entity are and their corresponding scientific names in scientific name library, so as to enhance the semantic features of the mention phrases in the sentences.

Many entities in Tibet animal husbandry dataset have the phenomenon of the multiple mentions, the names of some entities are neither abbreviations nor abbreviations. For example, "artemisia annua" has other mentions such as "mugwort", "zebra bile" and "twelve sisters", and these completely different mentions pose a challenge to the task of the coreference resolution. In the next step, we intend to fuse the entity description text with the entity mention, and make the semantic features of the mention closer to its scientific name features through fusion representation, so as to improve the coreference resolution ability of the model. Secondly, we use simple contrastive learning of sentence embedding (SimCSE) model to embed representation text, which is the best text representation model at present. The SimCSE model is used to further improve the contextual semantic representation of the mention, this will certainly help irregular mentions such as non-abbreviations and abbreviations to match with their scientific names. Thirdly, we will try to train named entity recognition and reference resolution tasks in the same model to reduce error propagation of neural network.

REFERENCES

- [1] V. Dobrovolskii, "Word-level coreference resolution," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Punta Cana, DR, 2021, pp. 7670–7675.
- [2] J. J. Lin, Y. Z. Zhao, C. F. Liu, and T. Q. and Gao, "Entity coreference resolution for syllabus via graph neural network," in *Proc. FICC*, Mar. 2021, pp. 396–403, doi: [10.1007/978-981-16-1160-5_31](https://doi.org/10.1007/978-981-16-1160-5_31).
- [3] H. Lee, M. Surdeanu, and D. Jurafsky, "A scaffolding approach to coreference resolution integrating statistical and rule-based models," *Natural Lang. Eng.*, vol. 23, no. 5, pp. 733–762, 2017.
- [4] V. Ng, "Supervised noun phrase coreference research: The first fifteen years," in *Proc. ACL*, Uppsala, Sweden, 2010, pp. 1396–1411.
- [5] T. L. Thi, T. P. Thi, and T. Q. Thanh, "Machine learning using context vectors for object coreference resolution," *Computing*, vol. 18, pp. 1–20, Jan. 2021, doi: [10.1007/s00607-021-00902-4](https://doi.org/10.1007/s00607-021-00902-4).
- [6] M. Yuan, P. Xia, C. May, B. Van Durme, and J. Boyd-Graber, "Adapting coreference resolution models through active learning," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 7533–7549.
- [7] Y. Wang, Y. Shen, and H. Jin, "An end-to-end actor-critic-based neural coreference resolution system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 7848–7852.
- [8] C. Park, J. Shin, S. Park, J. Lim, and C. Lee, "Fast End-to-end coreference resolution for Korean," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, Stroudsburg, PA, USA, 2020, pp. 2610–2624.
- [9] R. Sukthankar, S. Poria, E. Cambria, and R. Thirunavukarasu, "Anaphora and coreference resolution: A review," *Inf. Fusion*, vol. 59, pp. 139–162, Jul. 2020, doi: [10.1016/j.inffus.2020.01.010](https://doi.org/10.1016/j.inffus.2020.01.010).
- [10] R. Guarasci, A. Minutolo, E. Damiano, G. De Pietro, H. Fujita, and M. Esposito, "ELECTRA for neural coreference resolution in Italian," *IEEE Access*, vol. 9, pp. 115643–115654, 2021. [Online]. Available: <https://ieeexplore.ieee.org/document/9514855>
- [11] B. Z. Li, G. Stanovsky, and L. Zettlemoyer, "Active learning for coreference resolution using discrete annotation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8320–8331. [Online]. Available: <https://aclanthology.org/2020.acl-main.738.pdf>
- [12] A. Zhukova, F. Hamborg, K. Donnay, and B. Gipp, "XCoref: Cross-document coreference resolution in the wild," in *Proc. 17th Int. Conf. Inf.*, Feb. 2022, pp. 272–291, doi: [10.1007/978-3-030-96957-8_25](https://doi.org/10.1007/978-3-030-96957-8_25).
- [13] I. Azerkovich, "Using semantic information for coreference resolution with neural networks in Russian," in *Proc. AIST*, Kazan, Russia, 2020, pp. 85–93.
- [14] A. Sboev, R. Rybka, and A. Gryaznov, "Deep neural networks ensemble with word vector representation models to resolve coreference resolution in Russian," in *Advanced Technologies in Robotics and Intelligent Systems*, vol. 80. Berlin, Germany: Springer, Jan. 2020, pp. 35–44, doi: [10.1007/978-3-030-33491-8_4](https://doi.org/10.1007/978-3-030-33491-8_4).
- [15] P. Kopyt, B. Salski, P. Zagrajek, D. Janczak, M. Sloma, M. Jakubowska, M. Olszewska-Placha, and W. Gwarek, "Electric properties of graphene-based conductive layers from DC up to terahertz range," *IEEE Trans. Terahertz Sci. Technol.*, vol. 6, no. 3, pp. 480–490, May 2016, doi: [10.1109/TTHZ.2016.2544142](https://doi.org/10.1109/TTHZ.2016.2544142).
- [16] R. Nair, V. N. V. Prasad, A. Sreenadh, and J. J. Nair, "Coreference resolution for ambiguous pronoun with BERT and MLP," in *Proc. Int. Conf. Adv. Comput. Commun. (ICACC)*, Kochi, India, Oct. 2021, pp. 1–5.
- [17] L. Liu, Z. G. Huan, G. Q. Jiang, M. Liu, and K. Ding, "Employing gated mechanism to incorporate symbolic features into Chinese event coreference resolution," in *Proc. CECIT*, Singapore, 2021, pp. 549–554.
- [18] R. Zhang, C. N. dos Santos, M. Yasunaga, B. Xiang, and D. Radev, "Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, 2018, pp. 102–107.
- [19] H. Y. Cheng, P. F. Li, and Q. M. Zhu, "Event coreference resolution method based on attention mechanism," *Comput. Sci.*, vol. 46, no. 9, pp. 201–205, Sep. 2019.
- [20] J. Ma, J. Liu, Y. Li, X. Hu, Y. Pan, S. Sun, and Q. Lin, "Jointly optimized neural coreference resolution with mutual attention," in *Proc. 13th Int. Conf. Web Search Data Mining*, Houston, TX, USA, Jan. 2020, pp. 402–410.
- [21] F. Jiang and T. Cohn, "Incorporating syntax and semantics in coreference resolution with heterogeneous graph attention network," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 1584–1591. [Online]. Available: <https://github.com/Fantabulous-J/coref-HGAT>
- [22] X.-F. Wang, L. Wang, A. Hawbani, and F.-Y. Miao, "Aspect level sentiment classification with memory network using word sentiment vectors and a new attention mechanism AM-PPOSC," in *Proc. IEEE 20th Int. Conf. High Perform. Comput. Commun., IEEE 16th Int. Conf. Smart City; IEEE 4th Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Exeter, U.K., Jun. 2018, pp. 1058–1063.
- [23] X. Guo, H. Zhang, H. Yang, L. Xu, and Z. Ye, "A single attention-based combination of CNN and RNN for relation classification," *IEEE Access*, vol. 7, pp. 12467–12475, 2019.
- [24] H. Wei, Z. Li, C. Zhang, and H. Ma, "The synergy of double attention: Combine sentence-level and word-level attention for image captioning," *Comput. Vis. Image Understand.*, vol. 201, Dec. 2020, Art. no. 103068.
- [25] D. P. Kingma and J. Ba, "ADAM: Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–5.
- [26] K. Ming, "Chinese coreference resolution via bidirectional LSTMs using word and token level representations," in *Proc. 16th Int. Conf. Comput. Intell. Secur. (CIS)*, Nanning, China, Nov. 2020, pp. 73–76.
- [27] J. Yang, S. Li, S. Gao, and J. Guo, "CorefDPR: A joint model for coreference resolution and dropped pronoun recovery in Chinese conversations," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 571–581, 2022, doi: [10.1109/TASLP.2022.3140545](https://doi.org/10.1109/TASLP.2022.3140545).
- [28] K. Lee, L. He, and L. Zettlemoyer, "Higher-order coreference resolution with coarse-to-fine inference," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, New Orleans, LA, USA, 2018, pp. 687–692.
- [29] W. Wu, F. Wang, A. Yuan, F. Wu, and J. W. Li, "CorefQA: Coreference resolution as query-based span prediction," in *Proc. 58th ACL*, 2020, pp. 6953–6963. [Online]. Available: <https://aclanthology.org/2020.acl-main.622.pdf>



BIANBADROMA was born in Tibet, China, in 1989. She received the B.M. degree in industry and business administration from the Central University of Finance and Economics, Beijing, China, in 2012. She is currently pursuing the master's degree.

Since graduating from undergraduate course, she has been working in big data and artificial intelligence technology at Tibet Net-Cloud Sci-Tech Company Ltd., Lhasa, Tibet, where she is currently the Technical Director. Her current research interests include big data and artificial intelligence technology.



NGODRUP was born in Tibet, China, in 1962. He received the B.S. degree in mathematics major from Tibet University, Lhasa, Tibet, in 1986, and the M.S. degree in computer software and theory from the University of Bergen, Bergen, Norway, in 2000.

From 2000 to 2015, he devoted himself to Tibetan information processing research at Tibet University. He was a Professor of computer science and technology, a Doctoral Supervisor of computational linguistics, and the Director of the Tibetan Information Technology Research Center, Tibet University. Since 2016, he has been a Professor and a Master's Tutor with Xizang Minzu University. He is currently the Head of the Innovation Team, Ministry of Education, Tibetan Information Processing Technology; and the Director of the National Local Joint Engineering Research Center, Tibetan Information Technology. He has completed more than 20 national, provincial, and ministerial scientific research and teaching projects. He has published five academic monographs, including "Tibetan Computational Linguistics" and "Tibetan Pattern Recognition Technology and Engineering Practice." He has published more than 30 academic papers, such as "A Study on Tibetan Script Encoding in the UCS" and "Study on Printed Tibetan Character Recognition," and been granted more than 20 national invention patents and software copyrights and ten national standards. His research interests include Tibetan information processing technology and intelligent speech technology.

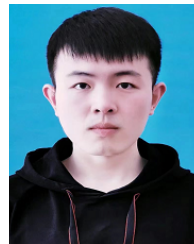
Prof. Ngodrup is a member of the Management Department of Science and Technology Committee of the 7th Ministry of Education and the Executive Director of the Chinese Information Society of China. He has won honorary titles, such as the Second Prize of National Science and Technology Progress, the First Prize of China Standard Innovation Contribution, the First Prize of Tibet Autonomous Region Science and Technology, the National Outstanding Science and Technology Worker, and the Special Prize of Invention and Entrepreneurship Award.



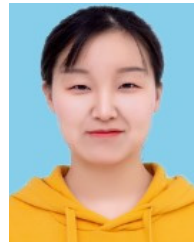
ERPING ZHAO was born in Binxian, Shaanxi, China, in 1976. He received the B.E. degree in computer application and the M.S. degree in software engineering from Xidian University, Xi'an, Shaanxi, in 1999 and 2006, respectively.

From September 2016 to July 2017, he was a Visiting Scholar at the Renmin University of China, where he conducted the researches on the knowledge graph. From 2006 to 2012, he was a Lecturer with the School of Information Engineering, Xizang Minzu University, where he was an Associate Professor, from 2013 to 2021, and has been a Professor, since 2022. His research interests include big data analysis and application, and knowledge graph.

Prof. Zhao was a Senior Member of the China Computer Federation (CCF) and an Executive Member of the Information System Professional Committee, CCF.



YUHAO WANG was born in 1996. He is currently pursuing the master's degree. His current research interests include knowledge graph completion, representation learning, and natural language processing.



YAKUN ZHANG was born in Shanxi, China, in 1997. She is currently pursuing the master's degree. Her current research interests include named entity recognition and natural language processing.

...