Tech Science Press

Hybrid Neural Network for Automatic Recovery of Elliptical Chinese Quantity Noun Phrases

Hanyu Shi[1], Weiguang Qu[1,2,*], Tingxin Wei[2,3], Junsheng Zhou[1], Yunfei Long[4], Yanhui Gu[1] and Bin Li[2]

[1]School of Computer Science and Technology, Nanjing Normal University, Nanjing, 210023, China
[2]School of Chinese Language and Literature, Nanjing Normal University, Nanjing, 210097, China
[3]International College for Chinese Studies, Nanjing Normal University, Nanjing, 210097, China
[4]School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK
[*]Corresponding Author: Weiguang Qu. Email: wgqu_nj@163.com
Received: 15 April 2021; Accepted: 04 June 2021

Abstract: In Mandarin Chinese, when the noun head appears in the context, a quantity noun phrase can be reduced to a quantity phrase with the noun head omitted. This phrase structure is called elliptical quantity noun phrase. The automatic recovery of elliptical quantity noun phrase is crucial in syntactic parsing, semantic representation and other downstream tasks. In this paper, we propose a hybrid neural network model to identify the semantic category for elliptical quantity noun phrases and realize the recovery of omitted semantics by supplementing concept categories. Firstly, we use BERT to generate character-level vectors. Secondly, Bi-LSTM is applied to capture the context information of each character and compress the input into the context memory history. Then CNN is utilized to capture the local semantics of n-grams with various granularities. Based on the Chinese Abstract Meaning Representation (CAMR) corpus and Xinhua News Agency corpus, we construct a hand-labeled elliptical quantity noun phrase dataset and carry out the semantic recovery of elliptical quantity noun phrase on this dataset. The experimental results show that our hybrid neural network model can effectively improve the performance of the semantic complement for the elliptical quantity noun phrases.

Keywords: Elliptical quantity noun phrase; semantic complement; neural network

1 Introduction

Ellipsis is a widespread linguistic phenomenon that aims to recover missing objects in the sentence based on the surrounding discourse. The human can quickly restore the omitted semantics based on context, however, it is a big challenge for machine understanding. The gap in syntax and the omission in semantics bring obstacles to many NLP tasks such as syntactic parsing, machine translation and information extraction. Therefore, ellipsis recovery has become one of the fundamental tasks in natural language processing.

Quantity noun phrase is one of the most common structures in Chinese, in which quantity phrase is used to modify and specify the noun head. However, in some contexts, the noun head is omitted, while the quantity phrase is used to represent the meaning of the whole quantity noun phrase. According to Dai et al. [1], the elliptical quantity phrase accounts for 38.7% of the omission in the corpus of Chinese Abstract Meaning Representation (CAMR). As the omitted noun bears the core meaning of the quantity noun phrase, it is necessary and critical to restore it for the downstream tasks of natural language understanding [2]. However, this research topic has been neglected by most of the research communities. To fill this gap, our research proposes a hybrid neural network model for automatically complement the elliptical quantity noun phrase in Chinese to fill the current research gap.

Quantity noun phrase, usually in the form of "numerical word + quantifier + noun", is a common type of phrase in modern Chinese, in which numerical words and quantifiers are the noun's modifiers. For example, in the case of "一*(one)*个*(unit)* 学生 *(student)*", the numeral "一*(one)*" and the quantifier "个*(unit)*" are the modifiers of the noun head "学生*(student)*". In actual language use, following the economic principle of pragmatics, if the head of a quantity noun phrase appears in the context as an antecedent, the noun head can be omitted and the quantity noun phrase is reduced to a quantity phrase and forms an elliptical quantity noun phrase. For example:

店家*(The clerk)*无奈*(had no choice)*，只好*(but to)*又*(again)*给*(for)*武松*(Wu Song)*筛酒*(pour spirits)*。武松*(Wu Song)*前后*(from beginning to end)*共*(altogether)*吃了*(drunk)*十八*(eighteen)*碗*(bowls)*[酒*(spirits)*]*。

The clerk had no choice but to pour spirits for Wu Song. Wu Song drunk eighteen bowls [of spirits] altogether.

In CAMR, the concept nodes of omitted arguments are allowed to be added, which makes the semantic representation of sentences more natural and comprehensive. As is shown in Fig. 1, CAMR recovers the noun head of the elliptical quantity noun phrase "十八碗*(eighteen bowls)*" by adding the concept *thing*.

```
武松前后共吃了十八碗。
Wu Song drunk eighteen bowls altogether.
        吃(had/ate)-01
                :arg0()         /person
                        :name()         武松(Wu Song)
                :mod()          前后(from beginning to end)
                :mod()          共(altogether)
                :aspect()       了(particle for past tense)
                :arg1()         /thing
                        :quant()        18(eighteen)
                        :cunit()        碗(bowl)
```

Figure 1: An example of concept addition of an elliptical quantity noun phrase

The quantity phrase "十八碗 *(eighteen bowls)*" is the elliptical form of the quantity noun phrase "十八碗[酒]*(eighteen bowls of spirits)*", as the noun head "酒 *(spirits)*" appears in the previous sentence. In our work, we realize semantic recovery by supplementing concept categories that are automatically distinguished by our model. Semantic recovery of elliptical quantity noun phrases will help the subsequent Chinese Abstract Meaning Representation (CAMR) parsing by improving the accuracy of the parser; hence, it is beneficial to subsequent work based on the CAMR such as machine translation [3,4], text summarization [5], event extraction [6] and other works.

AMR is an abstract meaning representation method which represents the semantics of a sentence with a single rooted directed acyclic graph [7]. What distinguishes it from the other sentential semantic representation methods is that the elliptical arguments are also represented with concept nodes in its semantic graph [8]. There are overall 109 concepts in its Name Entity List that can be used to represent elliptical arguments in AMR. Inspired by AMR (Abstract Meaning Representation) [9], we utilize concepts to complete the noun head of elliptical quantity noun phrases. This paper uses the concepts from Name Entity List of AMR to denote the omitted noun heads in elliptical quantity noun phrases [10]. In the case "武松前后共吃了十八碗。*(Wu Song drunk eighteen bowls [of spirits] altogether.)*", we use the concept "*thing*" to represent the head that is omitted by the quantity phrase "十八碗 *(eighteen bowls)*". Moreover, in our work, we build an elliptical quantity noun phrase dataset based on the Chinese AMR corpus [11] and Xinhua News Agency corpus [12].

Since both the characters and words in the elliptical quantity noun phrases are crucial to recover the omitted noun head, we focus on the information of n-grams with various granularities in the target sentence. We utilize a hybrid neural network to obtain sufficient semantic information from the elliptical quantity noun phrase and its context for the omission recovery. We aim to recover the omitted head of an elliptical quantity noun phrase automatically with a hybrid neural network model, combining bidirectional encoder representation transformer (BERT) [13], bidirectional long short-term memory network (Bi-LSTM) [14] and convolutional neural network (CNN) [15].

To be specific, firstly, we utilize BERT to generate character-level vectors. Secondly, we utilize Bi-LSTM to capture both the preceding and following context information of each character and compress the input into the context memory history. Then CNN is utilized to capture the local semantics of n-grams with various granularities. Finally, the semantic representation of elliptical quantity noun phrase and its context are both used to predict the probabilities of the concepts that can represent the elliptical component.

To my knowledge, this is the first time that use the deep learning method on the Chinese elliptical quantity noun phrase dataset and fill the gaps in the research on semantic complement of elliptical quantity noun phrase. The main contributions of this paper are as follows: 1. We build a hand-labeled dataset of elliptical quantity noun phrases. 2. The method based on the hybrid neural network model is first proposed to complement elliptical quantity noun phrases. We anticipate that our model could recover the omitted noun head of the elliptical quantity noun phrases in a high accuracy, and subsequently helps to improve the downstream tasks.

2 Related Work

In this paper, we utilize the neural network model to predict the concept for complement the elliptical quantity noun phrases. Our work is related to previous works on quantity noun phrase research, ellipsis recovery and the related knowledge of AMR.

2.1 Quantity Noun Phrase

For the semantic research of quantity noun phrase, the existing works mainly focus on the boundary recognition of quantity noun phrases, mostly by adopting rules-based and database-based methods to identify the structural features of quantity noun phrases. Zhang et al. [16] built a database that provides vocabulary knowledge and phrase structure knowledge for numerical phrase recognition and implemented the boundary recognition of "numerical words + quantifier". Xiong et al. [17] proposed a rule-based Chinese quantitative phrase recognition method based on Zhang's method without word segmentation. Fang et al. [18] proposed a backoff algorithm to get the quantifier-noun collocations that are not collected in dictionaries, which resulted in greatly improved recall in quantity noun phrase recognition. However, no researchers have paid attention to the semantic recovery of elliptical quantity noun phrases yet. Therefore, this paper aims to recover the omitted noun heads for elliptical quantity noun phrases to provide accurate and comprehensive sentential semantic representation for downstream semantic analysis and applications.

2.2 Ellipsis Recovery

In recent years, there has been a lot of work on semantic Chinese omission recovery. Shi et al. [19] explored a method to automatically complement the omitted head of elliptical DE phrases with a hybrid model, by combining densely connected Bi-LSTM and CNN(DC-Bi-LSTM). Zhang et al. [20] proposed a neural framework to uniformly integrate the two subtasks, VPE (Verb Phrase Ellipsis) detection and VPE resolution. For VPE detection, Zhang chose an SVM model with non-linear kernel function, a simple multilayer perception (MLP) and the Transformer. For VPE resolution, Zhang applied a MLP and the Transformer model, and finally a novel neural framework is proposed to integrate the two subtasks uniformly. Wu et al. [21] present CorefQA (Coreference Resolution as Query-based Span Prediction), an accurate and extensible approach for the coreference resolution task. Wu formulated the problem as a span prediction task, like in question answering: A query is generated for each candidate mention using its surrounding context, and a span prediction module is employed to extract the spans of the coreferences within the document using the generated query. Existing works to solve omission recovery are based on the specific phrase structure of the task, and might not suitable for all omission recovery tasks. We learn from its methods and research ideas, and carry out omission recovery work on the elliptical quantity noun phrases in this article.

2.3 Abstract Meaning Representation (AMR)

AMR [9] represents the semantics of sentences with a single rooted directed acyclic graph and allows sharing arguments. As it is more specific and accurate in semantic representation than other semantic representation methods, it has been applied to and proved effective in many NLP tasks such as event extraction [22], natural language modeling [23], text summarization [24], machine translation [25] and question answering [26]. Chinese AMR corpus (CAMR) was released in 2019 containing 10,419 Chinese sentences annotated with 109 concepts. We create our own dataset by

extracting the elliptical quantity noun phrases from CAMR corpus and Xinhua News Agency corpus [12].

By studying and analyzing the existing related work of omission recovery, we use neural network methods to implement the semantic completion for elliptical quantity noun phrases in this paper. We regard the head complement of elliptical quantity noun phrases as a classification task, and propose a hybrid model to determine the corresponding noun head for a concept.

3 Methodology

3.1 Model Overview

An elliptical quantity noun phrase in Chinese sentence can be denoted as: $S = \{s_1, s_2, \ldots, s_{y-1}, s_y^q, s_{y+1}^q, \ldots, s_{y+m}^q, s_{y+m+1}, \ldots s_n\}$, where s_i represents the character in the sentence, $1 \leq i \leq n$, $1 \leq y \leq n$, m is the length of elliptical quantity noun phrase. $\{s_1, \ldots, s_{y-1}\}$ represents the preceding context words of the quantity noun phrase, $\{s_y^q, s_{y+1}^q, \ldots, s_{y+m}^q\}$ represents the elliptical quantity noun phrase or quantity noun phrase, $\{s_{y+m+1}, \ldots, s_n\}$ represents the following context words. This paper aims to select the correct concept to complement the omitted head of the elliptical quantity noun phrase. We use the powerful BERT pre-training model, which can generate character-level embedding vectors combined with contextual information. Subsequently, Bi-LSTM is used to further generate a textual representation based on contextual semantics. Then, the CNN model is used to extract feature in sentences. Therefore, after Bi-LSTM semantic encoding, we use the feature extraction module to obtain various granularities features.

The overview of the model framework proposed in this paper is shown in Fig. 2: Overall, the model has four principal modules:
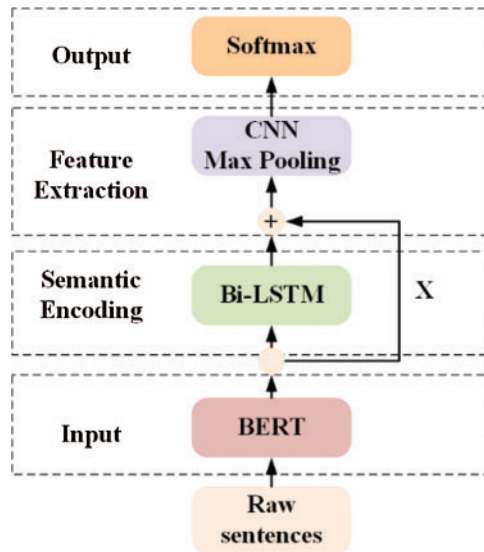
Figure 2: The illustration of model architecture

- Input Module: Input the raw corpus and generate character-level vectors through BERT pre-training.

- Semantic Encoding Module: Bi-LSTM model is utilized to obtain the semantic information of each word in the elliptical quantity noun phrase and its context.
- Feature Extraction Module: CNN model and Max pooling are used to capture semantics of n-grams with various granularities.
- Output Module: Generate predicted probability distribution over all concepts.

We will introduce the details of above four modules in the rest of this section.

3.2 Input Module

In the input module, we feed the data set of elliptical quantity noun phrases into the BERT pre-training module by character to obtain character vectors combined with contextual semantics. We take the output of this module based on the character-level vector of BERT as the input of the semantic encoding module.

BERT is a pre-trained language model proposed in 2018. Due to its powerful text representation capability, the model provides strong support in many different natural language processing tasks [27], especially in semantic analysis and semantic similarity computing. BERT pre-trains a general "language understanding" model in an unsupervised manner on a very large-scale corpus. When it is applied to other downstream tasks, it only needs to fine-tune the parameters of the network according to the specific task requirements [28].

The sub-structure of BERT is the encoder block of Transformer [29]. BERT abandons the recurrent network structure of RNN, regards the Transformer encoder as the main structure of the model, and uses the self-attention mechanism to model sentences. Its model structure is shown in Fig. 3.
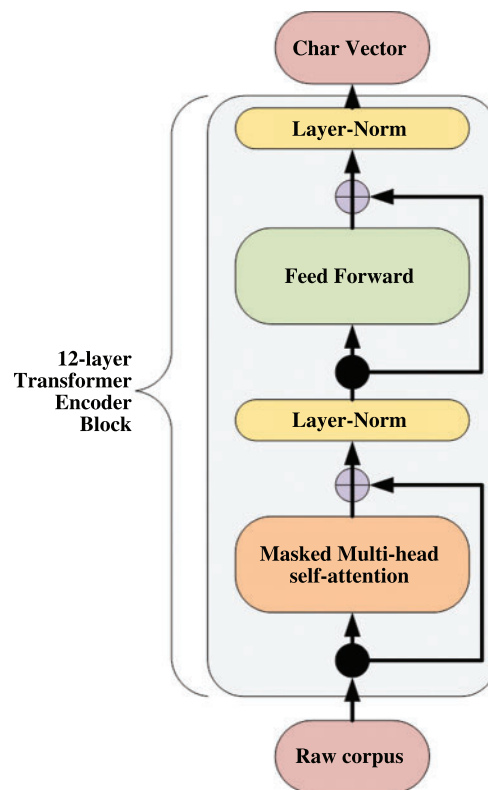

Figure 3: BERT model structure

In the model, multi-head self-attention is calculated in the same way with the conventional attention mechanism, while Layer-Norm is that the upper layer normalizes the residual connection. And Feed Forward denotes a two-layer linear transformation calculation. The entire BERT model consists of 12 layers of Transformer.

As shown in Eq. (1), in order to obtain the character-level embedding vector pre-trained by the corpus, the original corpus $S = \{s_1, s_2, \ldots, s_{y-1}, s_y^q, s_{y+1}^q, \ldots, s_{y+m}^q, vs_{y+m+1}, \ldots, s_n\}$ is processed into token embedding, segment embedding, position embedding. These 3 parts are used as input to the BERT model.

$$X = token(S) + segment(S) + position(S) \tag{1}$$

In Eq. (1), two special tokens need to be embedded, where the [CLS] is the tag at the beginning of each sentence and the [SEP] is the tag at the end of the sentence. Since the elliptical quantity noun phrases always appear in one sentence there is no other sentences that [CLS] is used to tag. For input, we fill the segment embedding with 0 s. Besides, position embedding is mainly used to encode the sequence order. The BERT model jointly adjusts the internal two-way Transformer encoder and uses the self-attention mechanism to learn the contribution of the remaining characters in the context to the current character, thereby enhancing contextual semantic information acquisition. Finally, BERT model generates character-level embedding vectors $X = \{x_1, x_2, \ldots, x_{y-1}, x_y^q, x_{y+1}^q, \ldots, x_{y+m}^q, x_{y+m+1}, \ldots x_n\}$ based on the current context.

3.3 Semantic Encoding Module

Semantic encoding module obtains the character-level vector X from the previous module as input. This module uses the two-layer Bi-LSTM model to perform semantic encoding and passes the Bi-LSTM hidden layer value h as the output to the local feature extraction module.

Long short-term memory neural network (LSTM) is a specific type of recurrent neural network (RNN) [30] that introduces memory mechanism and forgetting mechanism for neurons in hidden layers [31]. It effectively addresses the issue of long-distance dependence in RNN and avoids the gradient exploding or vanishing in the stage of error backpropagation calculation caused by long input sequence.

As shown in Fig. 4, each memory element of LSTM contains one or more memory blocks and three adaptive multiplication gates, i.e., Input Gate, Output Gate and Forget Gate, through which information can be saved and controlled. LSTM computes the hidden states $h_t = \{h_1, h_2, \ldots, h_t\}$ by iterating the following equation:

$$h_t = lstm(h_{t-1}, X_t) \tag{2}$$

At time step t, the memory c_t and the hidden state h_t are updated with the following equations:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ g \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} W_{lstm} \cdot [h_{t-1}; X_t] \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g \tag{4}$$

$$h_t = o_t \odot \tanh(c_t) \tag{5}$$

where X_t is the input at current time-step, i, f, o, g are respectively input gate, forget gate, output gate and new candidate memory state, σ and *tanh* are respectively sigmoid and hyperbolic tangent activation function, and $\odot$ denotes element-wise multiplication.
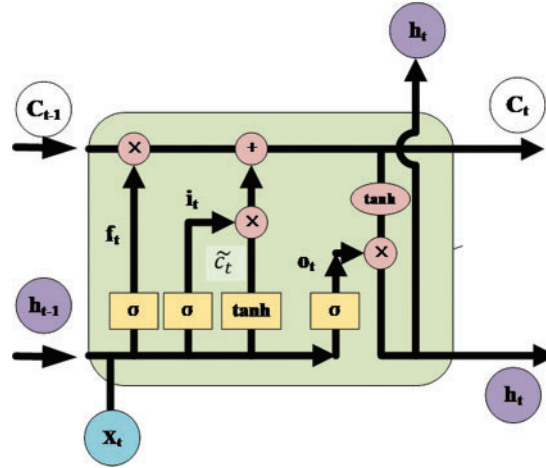

Figure 4: The Bi-LSTM architecture

However, the LSTM model, as a unidirectional neural network, is unable to learn the context information sufficiently, so Bi-LSTM is developed to improve semantic representation. In our work, we feed the character vector $X = \{x_1, x_2, \ldots, x_{y-1}, x_y^q, x_{y+1}^q, \ldots, x_{y+m}^q, x_{y+m+1}, \ldots x_n\}$ pretrained by BERT to the two-layer Bi-LSTM. The first layer Bi-LSTM output is fed to the second layer Bi-LSTM as input, and the second layer Bi-LSTM output is the concatenation of the last unit output of the forward and the backward propagation. The final output h is obtained through stacked Bi-LSTM layers:

$$h = [\overleftarrow{h'_1}, \overrightarrow{h'_n}] \tag{6}$$

The architecture of the proposed "BERT + Bi-LSTM" model is shown in Fig. 5.

3.4 Local Feature Extraction Module

Local feature extraction module combines the output of the semantic encoding module and the output of the BERT pre-training layer as the input of this module. This module feeds the input vector through the convolution neural network and max-pooling feature extraction to the new output module.

CNN is an effective model for extracting semantic features and capturing salient features in a flat structure [32]. The reason is that they are capable of capturing local semantics of n-grams with various granularities. While in the sentence containing an elliptical quantity phrase, the omitted noun head usually can be inferred by the context as the words in the context are informative. Therefore, for the feature extraction module, we use CNN and Max pooling layer to extract features. Meanwhile, in order to obtain more effective information and avoid the loss of semantic information during semantic encoding, we concatenate the character vectors obtained

by BERT pre-training with the semantic vectors generated by Bi-LSTM and input them into the feature extraction module.

$$K = [X, h] \tag{7}$$

We use a convolution filter $F \in R^{\omega \times d}$ to obtain the feature map $Y \in R^{n-\omega+1}$, the j-th element y_j is given by:

$$y_j = f(W \cdot K_{j:j+\omega-1} + b) \tag{8}$$

where f is the *Relu* activation function, W is the weight matrix of convolution filter, b is a bias, ω is the length of the filter, and d is the dimension of word vector. The convolutional layer uses multiple filters in parallel to obtain feature maps. We also can use convolutional filters of various length to extract different feature information.

The feature maps are fed to the max-pooling layer to obtain the most salient information to form the feature vectors that have been proved effective in improving the model performance.
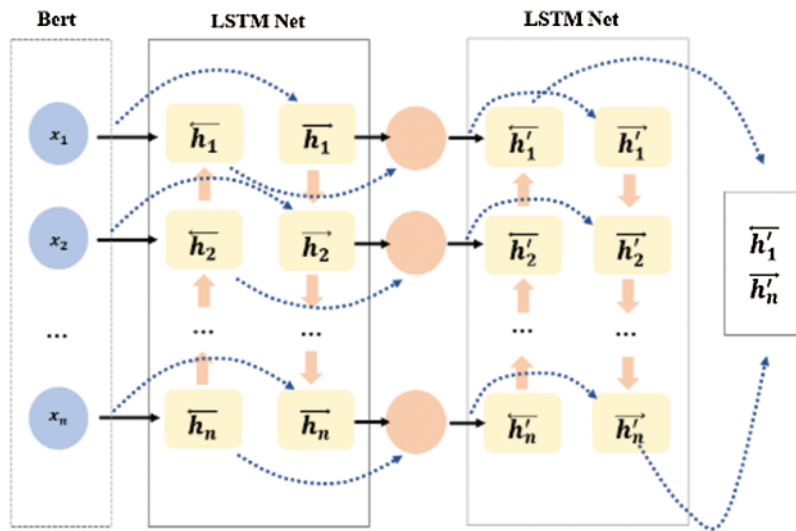

Figure 5: The illustration of "BERT + Bi-LSTM" architecture

3.5 Output Module

The output module uses the output of the previous feature extraction module to obtain the probability distribution of the concept category through the classification function.

With the semantic encoding module and the feature extraction module, we obtain the feature representation of elliptical quantity noun phrase and its context. Then we feed the output of local feature extraction module into *softmax* function to classify the omitted semantic categories of elliptical quantity noun phrase.

4 Experiments

The experiment in this paper utilizes a hybrid neural network model to recover elliptical quantity noun phrases and compares multiple sets of experiments to show the improvement of

our model. We first compare and analyze the experimental results, then discuss some error cases, and propose the feasible future improvements in this section.

4.1 Dataset

As there is no such dataset ready for use, we build our own elliptical quantity noun phrase dataset by extracting all the head omitted quantity noun phrases from the CAMR corpus and Xinhua News Agency corpus. In this dataset, we extract 838 elliptical quantity noun phrases from the CAMR corpus and 676 elliptical quantity noun phrases from the Xinhua News Agency corpus. We observe that in the corpus the omitted head concepts of the elliptical quantity noun phrases are only distributed on few categories such as *thing* and *person*, accounting for 85% of the overall elliptical quantity noun phrases. Tab. 1 shows the distribution of the elliptical quantity noun phrases and their concept categories in the training and test sets.

Table 1: Concept distribution the experimental dataset

Concept	Train	Test	All
Thing	783	241	1024
Person	199	79	278
Organization	66	19	85
Location	34	14	48
Animal	53	26	79

4.2 Experiment Settings

The experiment settings of the combined neural network based on our Hybrid Neural Network are shown in Tab. 2.

Table 2: Settings of hyperparameters

Hyperparameter	Parameter value
Epoch	100
Batch_size	5
BERT_hidden_size	768
Bi-LSTM_hidden_size	256
Bi-LSTM_layer_number	2
Filter_size	2, 3, 4
Num_filter	100
Dropout	0.1
Learning_rate	0.005
Optimization	Adam

4.3 Evaluation Metrics

We use accuracy and Macro-F1 score to measure the performance of our proposed method. Macro-F1 gives equal weight to each class label which is suitable for classification task with unbalanced categories [33].

4.4 Results

We conduct comparative experiments to prove the effectiveness of our proposed method, the results are shown in Tab. 3. CharVec means that we adopt 300-dimensional character embedding vector obtained from the Baidu Encyclopedia corpus trained by Skip-Gram with Negative Sampling (SGNS) [34] method. Bi-LSTM means a bidirectional LSTM network is used. ATT refers to attention mechanism for extracting features. CNN + Max pooling means using convolutional neural network and max-pooling to extract features.

Table 3: Performance of different methods

#	Method	P%	R%	F1%	Accuracy %
1	CharVec + Bi-LSTM	56.51	41.87	48.10	70.98
2	CharVec + Bi-LSTM + ATT	56.86	61.49	59.08	78.63
3	CharVec + CNN + Max pooling	76.47	67.12	71.49	82.32
4	CharVec + Bi-LSTM + CNN + Max pooling	79.43	65.64	71.89	83.38
5	BERT + Bi-LSTM + CNN + Max pooling	**89.32**	**86.48**	**87.88**	**92.08**

As shown in line 1 of Tab. 3, with the conventional character vector CharVec and Bi-LSTM model, we obtain an accuracy of 70.98%, which illustrates the effectiveness of baseline model complement the elliptical quantity noun phrase in this paper. When extra context features are incorporated, the performance is improved accordingly. Comparing method 4 with method 5, we find that the syntactic information provided through BERT pre-trained character vector plays a vital role in improving our proposed model. It can help locate important characters in the elliptical quantity noun phrase, capturing better contextual information in the following modules. Comparing method 2 with method 4, we can see that the CNN and Max pooling feature extraction methods perform better than the attention-based mechanism in this paper. That is because in the sentence of the elliptical quantity noun phrase, the words in the context are crucial for people to infer what is omitted. Based on this observation, we use convolutional neural networks to obtain contextual semantics of various granularities for feature extraction. As is shown in line 2 and 3, CNN and the max-pooling model obtain better results than Bi-LSTM+ATT, which also proves the effectiveness of CNN in extracting features.

Despite the achievement in complement, we notice that there exists unbalance among the categories of the concepts that are omitted in language, Tab. 4 shows the recognition results of models for each type of concept. The numbering sequence in the Tab. 4 is the same as in Tab. 3.

One can notice that the poor performance of location and animal do not affect the overall performance obviously, due to the small amount of data of these concepts. In addition to continuously improving the model to advance the experimental results, we can also use the method of expanding our experimental data set. More training corpus makes further improvements in the characteristics of model learning.

Table 4: Experiment results on different concepts

#	Thing			Person			Organization			Location			Animal		
	P%	R%	F1%	P%	R%	F1%	P%	R%	F1%	P%	R%	F1%	P%	R%	F1%
1	75.59	93.78	83.70	58.70	34.18	43.20	48.28	73.68	58.33	0	0	0	100.0	7.69	14.29
2	88.79	85.48	87.10	60.40	77.22	67.78	66.67	94.74	78.26	0	0	0	68.42	50.00	57.78
3	89.34	90.46	89.90	67.82	74.68	71.08	80.00	84.21	82.05	80.00	28.57	42.11	65.22	57.69	61.22
4	84.93	95.85	90.06	80.00	65.82	72.22	93.33	73.68	82.35	66.67	42.86	52.17	72.22	50.00	59.09
5	**95.44**	**95.44**	**95.44**	**83.33**	**88.61**	**85.89**	**100.0**	**100.0**	**100.0**	**76.92**	**71.43**	**74.07**	**90.91**	**76.92**	**83.33**

4.5 Case Study

Most of the concept complement errors occur in cross-sentences. We use the following sentence as an example and analyze it.

For example, "到 *(By)* 1997年 *(1997 year)* 年底 *(end of year)*, 全省 *(province)* 小城镇 *(small towns)* 达 *(had)* 830个 *(830)*。*(By the end of 1997, there were 830 small towns in the province.)*

而 *(But)*其中 *(among them)*, 面积 *(area)* 超过 *(more than)* 2平方公里 *(2 square kilometers)*的 [小镇 *(towns)*]只 *(only)* 有 *(had)* 213个 *(213)*[小镇 *(towns)*], 而且 *(and)* 布局 *(layout)* 分散 *(was scattered)*, 集聚 *(agglomeration)* 功能 *(function)* 差 *(was poor)*。*(Among them, only 213 [towns] had an area of more than 2 square kilometers, and the layout was scattered and the agglomeration function was poor.)*"

In this example, elliptical quantity noun phrase "213个 *(213)*" should be recognized as the *location* concept to complete the noun head of elliptical quantity noun phrase. Since our work uses one sentence as a unit for ellipsis recovery, the phenomenon of cross-sentence omission cannot be recovered well. This problem caused our model fail in obtaining information of "small towns" in the example sentence, and the elliptical quantity noun phrase "213" was classified into the category of *thing* concept. Therefore, in the later work, we will introduce the cross-sentence semantic information into the semantic complement task of elliptical quantity noun phrases, while expanding the data set to balance the elliptical quantity noun phrases of each concept category.

4.6 Ablation Experiment

The ablation experiments are implemented to evaluate the performance of each module. As shown in Tab. 5, we conduct four sets of experiments to compare the effects of the modules.

Table 5: Ablation experiment

#	Method	P%	R%	F1%	Accuracy %
I	BERT + Bi-LSTM	89.24	79.28	83.97	87.53
II	BERT + CNN + Max pooling	88.94	82.29	84.00	89.45
III	CharVec + Bi-LSTM + CNN + Max pooling	79.43	65.64	71.89	83.38
IV	BERT + Bi-LSTM + CNN + Max pooling	**89.32**	**86.48**	**87.88**	**92.08**

Comparing experiment I with experiment IV, local feature extraction module can improve the experimental results by 4.55%. Comparing experiment II with experiment IV, the accuracy of the semantic encoding module for the overall recovery increased by 2.63%. In the input module, the BERT pre-training module has the best improvement. Comparing experiment III with experiment IV, with pre-training, the accuracy of the model increased by 8.7%. This also reminds us that the training effect of larger-scale corpus can lead to better deep learning models. Enlarging the scale of the corpus is an effective way to improve the experimental results.

5 Conclusion

In this paper, we aim to recover the elliptical quantity noun phrases by identifying the corresponding semantic categories automatically. We propose a hybrid model of BERT-Bi-LSTM-CNN, which utilizes BERT to obtain good representation, the Bi-LSTM model to encode long sentences and CNN to extract features. Experiments show that our model can effectively complement the omission of elliptical quantity noun phrases. Recovery of elliptical Chinese quantity noun phrases help the subsequent Chinese Abstract Meaning Representation (CAMR) parsing by improving the accuracy of the parser; hence, it is beneficial to downstream tasks based on the CAMR.

In the future, we will explore the following directions. First, we will develop better models to recognize the elliptical quantity noun phrases with more concrete concepts categories. Second, we can also incorporate comprehensive linguistic knowledge of the elliptical quantity noun phrases into the model construction to improve the performance.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. B. Dai, B. Q. Shi, B. Li and W. G. Qu, "An investigation of Chinese ellipsis and argument sharing based on AMR corpus," *Foreign Languages Research*, vol. 37, no. 2, pp. 16–23, 2020.

[2] F. Xu, X. Zhang, Z. Xin and A. Yang, "Investigation on the Chinese text sentiment analysis based on convolutional neural networks in deep learning," *Computers, Materials & Continua*, vol. 58, no. 3, pp. 697–709, 2019.

[3] J. Y. M. Zhang, J. Liu and X. Y. Lin, "Improve neural machine translation by building word vector with part of speech," *Journal on Artificial Intelligence*, vol. 2, no. 2, pp. 79–88, 2020.

[4] R. V. Noord and J. Bos, "Neural semantic parsing by character-based translation: Experiments with abstract meaning representations," *Computational Linguistics in the Netherlands Journal*, vol. 7, pp. 93–108, 2017.

[5]   F. Liu, J. Flanigan, S. Thomson, N. M. Sadeh and N. A. Smith, "Toward abstractive summarization using semantic representations," in *Proc. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, USA, pp. 1077–1086, 2015.

[6]   S. Rao, D. Marcu, K. Knight and H. Daumé, "Biomedical event extraction using abstract meaning representation," in *Proc. of Biomedical Natural Language Processing*, Vancouver, Canada, pp. 126–135, 2017.

[7]   W. G. Qu, J. S. Zhou, X. D. Wu, R. B. Dai, M. Gu *et al.*, "Survey on abstract meaning representation," *Data Acquisition & Processing*, vol. 32, no. 1, pp. 26–36, 2017.

[8]   B. Li, Y. Wen, L. J. Bu, W. G. Qu and N. W. Xue, "A comparative analysis of the AMR graphs between English and Chinese corpus of the little prince," *Chinese Information Processing*, vol. 31, no. 1, pp. 50–57 + 74, 2017.

[9]   L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt *et al.*, "Abstract meaning representation for sembanking," in *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria, pp. 178–186, 2013.

[10] T. X. Wei, W. L. Bai and W. G. Qu, "Sense prediction for Chinese OOV based on word embedding and semantic knowledge," *Data Analysis and Knowledge Discovery*, vol. 4, no. 6, pp. 109–117, 2020.

[11] B. Li, Y. Wen, L. Song, L. J. Bu, W. G. Qu *et al.*, "Construction of Chinese abstract meaning representation corpus with concept-to-word alignment," *Chinese Information Processing*, vol. 31, no. 6, pp. 93–102, 2017.

[12] Web Download Philadelphia: Linguistic Data Consortium, "Chinese gigaword second edition LDC2005T14," 2005. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2005T14.

[13] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, pp. 4171–4186, 2019.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1746–1751, 2014.

[16] L. Zhang, W. Xiong, Y. Li and Y. Liu, "Recognize modern Chinese quantifier phrases based on a knowledge-database," in *Proc. of the 7th Int. Conf. on Chinese Information Processing*, Wuhan, Hubei, China, pp. 306–310, 2007.

[17] W. Xiong and L. Zhang, "A rule based and no word segmentation Chinese quantifier phrase chunking," in *Proc. of the 7th Int. Conf. on Chinese Information Processing*, Wuhan, Hubei, China, pp. 47–51, 2007.

[18] F. Fang and B. Li, "Corpus based investigation on MQN phrase," in *Proc. of the 3rd Sym. on Computational Linguistics for Students*, Shenyang, Liaoning, China, pp. 346–352, 2006.

[19] B. Q. Shi, W. G. Qu, R. B. Dai, B. Li, X. H. Yan *et al.*, "A general strategy for researches on Chinese "的(de)" structure based on neural network," *World Wide Web*, vol. 23, no. 6, pp. 2979–3000, 2020.

[20] W. Zhang, Y. Zhang, Y. Liu, D. Di and T. Liu, "A neural network approach to verb phrase ellipsis resolution," in *Proc. of the Thirty-Third Conf. on Artificial Intelligence*, Honolulu, Hawaii, pp. 7468–7475, 2019.

[21] W. Wu, F. Wang, A. Yuan, F. Wu and J. Li, "CorefQA: Coreference resolution as query-based span prediction," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington, USA, pp. 6953–6963, 2020.

[22] S. Rao, D. Marcu, K. Knight and H. Daumé III, "Biomedical event extraction using convolutional neural networks and dependency parsing," in *Proc. of Biomedical Natural Language Processing*, Melbourne, Australia, pp. 98–108, 2018.

[23] S. Garg, G. V. Steeg and A. Galstyan, "Stochastic learning of nonstationary kernels for natural language modeling," *Computer Science*, 2019. [Online]. Available: https://arxiv.org/abs/1801.03911.

[24] S. Dohare and H. Karnick, "Text summarization using abstract meaning representation," *Computer Science*, 2017. [Online]. Available: https://arxiv.org/abs/1706.01678.

[25] B. K. Jones, J. Andreas, D. Bauer, K. M. Hermann and K. Knight, "Semantics-based machine translation with hyperedge replacement grammars," in *Proc. of 24th Int. Conf. on Computational Linguistics*, Mumbai, India, pp. 1359–1376, 2012.

[26] A. Mitra and C. Baral, "Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning," in *Proc. of the Thirtieth Conf. on Artificial Intelligence*, Phoenix, Arizona, pp. 2779–2785, 2016.

[27] P. J. Xu, C. Li, L. G. Zhang, F. Yang, J. Zheng *et al.*, "Underground disease detection based on cloud computing and attention region neural network," *Journal on Artificial Intelligence*, vol. 1, no. 1, pp. 9–18, 2019.

[28] B. Oh, H. Song, J. Kim, C. Park and Y. Kim, "Predicting concentration of PM10 using optimal parameters of deep neural network," *Intelligent Automation & Soft Computing*, vol. 25, no. 2, pp. 343–350, 2019.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.*, "Attention is all you need," in *Proc. of Annual Conf. on Neural Information Processing Systems*, Long Beach, CA, pp. 5998–6008, 2017.

[30] M. Sundermeyer, H. Ney and R. Schlüter, "From feedforward to recurrent LSTM neural networks for language modeling," *IEEE/ACM Transactions on Audio Speech, and Language Processing*, vol. 23, no. 3, pp. 517–529, 2015.

[31] G. Yang, J. Zeng, M. Yang, Y. Wei and X. Wang, "Ott messages modeling and classification based on recurrent neural networks," *Computers, Materials & Continua*, vol. 63, no. 2, pp. 769–785, 2020.

[32] Y. Liu, F. Wei, S. Li, H. Ji, M. Zhou *et al.*, "A dependency-based neural network for relation classification," in *Proc. of Association for Computational Linguistics*, Beijing, China, pp. 285–290, 2015.

[33] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[34] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu *et al.*, "Analogical reasoning on Chinese morphological and semantic relations," in *Proc. of Association for Computational Linguistics*, Melbourne, Australia, pp. 138–143, 2018.