Semantic-aware Chinese Zero Pronoun Resolution with Pre-trained Semantic Dependency Parser

Lanqiu Zhang
Beijing Language and Culture University
zhang_lanqiu@163.com

Zizhuo Shen
Beijing Language and Culture University
blcushzz@gmail.com

Yanqiu Shao✉
Beijing Language and Culture University
yqshao163@163.com

Abstract

Deep learning-based Chinese zero pronoun resolution model has achieved better performance than traditional machine learning-based model. However, the existing work related to Chinese zero pronoun resolution has not yet well integrated linguistic information into the deep learning-based Chinese zero pronoun resolution model.This paper adopts the idea based on the pre-trained model, and integrates the semantic representations in the pre-trained Chinese semantic dependency graph parser into the Chinese zero pronoun resolution model. The experimental results on OntoNotes-5.0 dataset show that our proposed Chinese zero pronoun resolution model with pre-trained Chinese semantic dependency parser improves the F-score by 0.4% compared with our baseline model, and obtains better results than other deep learning-based Chinese zero pronoun resolution models. In addition, we integrate the BERT representations into our model so that the performance of our model was improved by 0.7% compared with our baseline model.

1 Introduction

Chinese zero pronoun resolution is a special task of coreference resolution (Zhao and Ng, 2007). Its purpose is to find the real referent of the omitted parts with syntactic functions in the text. These omitted parts are usually called zero pronouns, and their real referents are called antecedents. Below is a sentence with zero pronouns:

[我]之前没有听说过[她]，[$*pro_1*$]听说[$*pro*_2$]是个有才华的美女。(I have not heard of [her] before, [$*pro_1*$] heard that [$*pro_2*$] is a talented beauty.)

In this example, the referent of zero pronoun $*pro_1*$ is “我/I”, and the referent of zero pronoun $*pro_2*$ is “她/her”. Since the zero pronoun is not a real word in the text, its resolution is much more difficult than that of the overt pronoun. The existence of zero pronouns poses challenges for machines to automatically understand text.

The existing Chinese zero pronoun resolution models with better performance usually adopt the method of deep learning (Chen and Ng, 2016);(Liu et al., 2017);(Yin et al., 2017);(Yin et al., 2018a);(Yin et al., 2018b). The deep learning-based methods can make the model automatically extract the task-related distributed representations through end-to-end training, thereby avoiding the problem that traditional machine learning-based methods rely heavily on artificially designed feature templates (Chen and Ng, 2016). However, it is difficult for deep learning-based models to encode effective syntactic, semantic and other linguistic information only through end-to-end training. Many deep learning-based Chinese zero pronoun resolution models still use syntactic features extracted from the syntactic parsing tree as a supplement to distributed representations.

Intuitively, semantic information as a higher level linguistic information is also very important to the Chinese zero pronoun resolution task, however few studies have attempted to integrate semantic information into the Chinese zero pronoun resolution model. Therefore, how to effectively integrate semantic information into the Chinese zero pronoun resolution model is a challenging problem. With the development of semantic parsing, the performance of some sentence-level semantic parsers have made

©2020 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 19th China National Conference on Computational Linguistics, pages 832-842, Hainan, China, October 31 - November 1, 2020.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

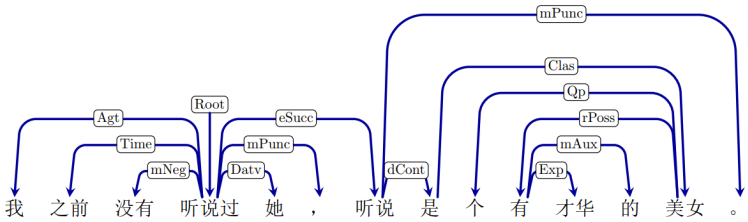

Figure 1: An example of a Chinese semantic dependency graph

remarkable progress, which provides opportunities for the application of sentence-level semantic parsing in other natural language processing tasks.

In this paper, we proposed a semantic-aware Chinese zero pronoun resolution model that integrates the semantic information from pre-trained Chinese semantic dependency graph parser. Chinese semantic dependency graph parsing (Che et al., 2016) is a semantic-level dependency parsing task, which is an extension of syntactic dependency parsing. Each node in the semantic dependency graph represents a word in the sentence, and the nodes are connected by directed edges with semantic relationship labels. Figure 1 is an example of a Chinese semantic dependency graph.

The realization of our model requires two stages. In the first stage, we use the Chinese semantic dependency graph parsing as a pre-training task to obtain a pre-trained semantic dependency graph parser. In the second stage, we feed the sentence which will be processed into the pre-trained semantic dependency graph parser to obtain the semantic-aware representations, and integrate these implicit semantic information into the Chinese zero pronoun resolution model.

We implement a attention-based Chinese zero pronoun resolution model as our baseline model. The experiments on OntoNotes-5.0 dataset show that our proposed Chinese zero pronoun resolution model with pre-trained Chinese semantic dependency parser improves the F-score by 0.4% compared with our baseline model, and obtains better results than other deep learning-based Chinese zero pronoun resolution models. In addition, we integrate the BERT representations into our model so that the performance of our model was improved by 0.7% compared with our baseline model.

2 Related Work

2.1 Zero Pronoun Resolution

Methods for solving Chinese zero pronoun resolution include rule-based methods, traditional machine learning-based methods, deep learning-based methods, etc. Converse (P and S, 2006) used Hobbs algorithm to traverse the syntactic tree of sentences to find the referent of zero pronoun. Zhao et al. (Zhao and Ng, 2007) designed more effective manual features for Chinese zero pronoun resolution task, and adopted a decision tree-based method to train supervised model. Kong et al. (Kong and Zhou, 2010) adopted a tree kernel-based method to model the syntax tree, so that the Chinese zero pronoun resolution model can make full use of the characteristics of the syntax tree. Chen et al. (Chen and Ng, 2016) designed a Chinese zero pronoun resolution model based on feed-forward neural network, and represented the zero-pronoun and candidate antecedent by combining manual feature vectors and word vectors, and obtained better performance than traditional machine learning-based methods. Yin et al. (Yin et al., 2017);(Yin et al., 2018a);(Yin et al., 2018b) designed a series of deep learning-based Chinese zero pronoun resolution model, which promoted the application of deep learning to Chinese zero pronoun resolution. Liu et al. (Liu et al., 2017) transformed the Chinese zero pronoun resolution task into the cloze-style reading comprehension task, and automatically constructed large-scale pseudo-data for the pre-training of their model.

2.2 Pre-training of Syntactic Dependency Parsing

Our method is similar to the method of pre-training of syntactic dependency parser, which has been successfully applied to some natural language processing tasks. Zhang et al. (Zhang et al., 2017) first

Proceedings of the 19th China National Conference on Computational Linguistics, pages 832-842, Hainan, China, October 31 - November 1, 2020.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

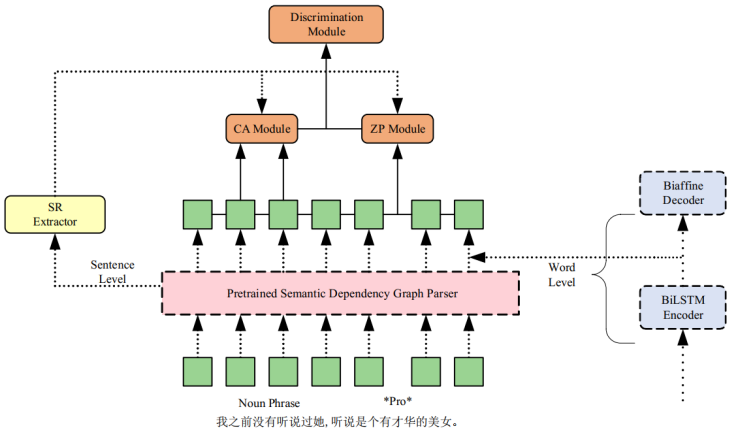

Figure 2: Chinese zero pronoun resolution model with pre-trained semantic dependency graph parser

proposed this method in the task of relation extraction. First, they trained the LSTM-based Biaffine syntactic dependency parser. Then, they extracted implicit syntactic representations from the LSTM layer of the well-trained syntactic dependency parser and integrated these representations into the relation extraction model. Guo et al. (Gao et al., 2017) and Yu et al. (Yu et al., 2018) used this method to integrate syntactic representations in the task of target-dependent sentiment analysis and discourse parsing respectively, and verified the effectiveness of this method in these tasks. Zhang et al. (Zhang et al., 2019) systematically studied the application of this method in the task of machine translation. Their experimental results show that this method obtains a more significant improvement than other methods such as Tree-Linearization and Tree-RNN in the task of machine translation. Jiang et al. (Jiang et al., 2020) applied this method to the task of Universal Conceptual Cognitive Annotation(UCCA) (Abend and Rappoport, 2013). Inspired by the method of integrating pre-trained information in ELMo (Peters et al., 2018), They made a weighted sum for the output of different LSTM layers of syntactic dependency parser. Their experimental results show that the method of fine-tuning pre-trained syntactic dependency parser improves the performance of UCCA model significantly.

3 Method

Given the success of the method of pre-training of syntactic dependency parser in some natural language processing tasks, we adopt a similar method to take the Chinese semantic graph dependency parsing as a pre-training task, and apply this method to Chinese zero pronoun resolution task.

Our proposed Chinese zero pronoun resolution model with pre-trained Chinese semantic dependency parser is composed of two parts, one is the pre-trained Chinese semantic dependency graph parser and the other is the Chinese zero pronoun resolution model. Specifically, The Chinese semantic dependency graph parser consists of two parts: BiLSTM-based encoder and Biaffine-based decoder. The Chinese zero pronoun resolution model consists of three parts: the zero pronoun module(ZP Module), the candidate antecedents module (CA Module) and the discrimination module. In addition, in order to obtain sentence-level semantic representations, we also used a CNN-based sentence representation extractor(SR Extractor).

For a sentence to be processed, the representations of each word will be feed into the pre-trained Chinese semantic dependency graph parser, so that each word can obtain the semantic-aware representations containing the information of semantic dependency graph. Then, the semantic-aware representations will be integrated into the Chinese zero pronoun resolution model to perform the subsequent processing. The overall architecture of our proposed model is shown in Figure 2:

3.1 Semantic Dependency Graph Parser

For the semantic dependency graph parser, we adopt 3-layer BiLSTM network and Biaffine network as encoder and decoder. The Biaffine-based parser has achieved the state of the art performance in some

tasks related to semantic dependency graph parsing. (Dozat and Manning, 2018);(Shen et al., 2019)

In the process of pre-training, we first use the concatenation of word vector, part of speech vector and character-level vector to represent a word. Then, we feed the word representations into the encoder to obtain the context-aware representations. Finally, we feed the context-aware representations of the word into the decoder to calculate the score of the dependency arc in the semantic dependency graph. The complete calculation process of the semantic dependency graph parser is shown in the following formulas:

$$w_t = [e_t^{(word)}; e_t^{(pos)}; e_t^{(char)}] \tag{1}$$

$$h_t = BiLSTM(w_t, h_{t-1}) \tag{2}$$

$$s_t^{(H,D)} = Biaffine(h_t^H, h_t^D) \tag{3}$$

where w_t means the word representations, h_t means the context-aware representations , $s_t^{(H,D)}$ means the score of the dependency arc, h_t^H and h_t^D mean context-aware representations of the head word and the dependent word respectively.

3.2 Zero Pronoun Module

According to the work of Yin et al (Yin et al., 2018b), we use BiLSTM network and self-attention mechanism to encode the preceding and following text of the zero pronoun. The purpose of using the self-attention mechanism is to obtain the attention weight distribution of the preceding and following word sequence. In this way, we can get the more powerful zero pronoun representations.

For a given anaphoric zero pronoun w_{zp}, we use $Context^{(pre)} = (w_1, w_2, \ldots, w_{zp-1})$ to denote the preceding word sequence of the zero pronoun, and use $Context^{(fol)} = (w_{zp+1}, w_{zp+2}, \ldots w_n)$ to denote the following word sequence of the zero pronoun. Each word w_t in the sentence is represented by the pre-trained word embedding.

In order to encode the contextual information of the word sequence, we first use two different 1-layer BiLSTM networks to separately process the the preceding word sequence and the following word sequence:

$$h_t^{(pre)} = BiLSTM^{(pre)}(w_t, h_{t-1}^{(pre)}) \tag{4}$$

$$h_t^{(fol)} = BiLSTM^{(fol)}(w_t, h_{t-1}^{(fol)}) \tag{5}$$

After that, we can obtain the preceding and following hidden vectors of the zero pronoun $h_t^{(pre)}$and $h_t^{(fol)}$ from the LSTM networks. we use $H^{(pre)}$ to denote the matrix which is concatenated by all preceding hidden vectors, and use $H^{(fol)}$ to denote the matrix which is concatenated by all following hidden vectors. where $H^{(pre)} \in \mathbb{R}^{n^{(pre)} \times d}$, $H^{(fol)} \in \mathbb{R}^{n^{(fol)} \times d}$, $n^{(pre)}$ and $n^{(fol)}$ means the number of words in the preceding and following word sequence respectively. d means the dimension of the hidden vectors.

The matrix $H^{(pre)}$ and $H^{(fol)}$ will be feed into the affine-based attention layers $Affine^{(pre)}$ and $Affine^{(fol)}$ to calculate the attention weight distribution of their associated sequences:

$$Affine(H) = Softmax(W_2 tanh(W_1 H^T)) \tag{6}$$

$$A^{(pre)} = Affine^{(pre)}(H^{(pre)}) \tag{7}$$

$$A^{(fol)} = Affine^{(fol)}(H^{(fol)}) \tag{8}$$

where $W_1 \in R^{h \times d}$, $W_2 \in R^{a \times h}$, $A^{(pre)} \in \mathbb{R}^{a \times n^{(pre)}}$, $A^{(fol)} \in \mathbb{R}^{a \times n^{(fol)}}$. It is worth explaining that a denotes the number of attention weight distributions. According to the work of Yin et al. (Yin et al., 2018b), we set the value of a to 2. Different attention weight distributions can capture different information, which further enhances the ability of the zero pronoun module.

Then, we can calculate the weighted sum of each row vector in the matrix by the following formula:

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

$$h_{zp}^{(pre)} = A^{(pre)} H^{(pre)} \tag{9}$$

$$h_{zp}^{(fol)} = A^{(fol)} H^{(fol)} \tag{10}$$

where $h_{zp}^{(pre)} \in R^{a \times d}$, $h_{zp}^{(fol)} \in R^{a \times d}$, If a is not equal to 1, We need to calculate the average of its row vectors.

At last, We take the concatenation of these two vectors as the final zero pronoun representations:

$$h_{zp} = [h_{zp}^{(pre)}; h_{zp}^{(fol)}] \tag{11}$$

3.3 Candidate Antecedents Module

When building the candidate antecedents module, we need to consider two types of the key information for the candidate antecedents. The first type of information is the context information of the candidate antecedents, and the second type of information is the interactive information between the zero pronoun and the candidate antecedents. Inspired by previous work (Lee et al., 2017), we use the context-aware boundary representations to capture context information and use attention mechanism to capture interactive information.

The candidate antecedent is usually a noun phrase composed of several words. So, we use $NP = (np_1, np_2, \ldots, np_n)$ to denote the set of all candidate antecedents for a given zero pronoun w_{zp} , and use $np_t = (w_i, w_2, \ldots, w_j)$ to denote a candidate antecedent within the set. First, we feed the pre-trained word vectors into the 1-layer BiLSTM network to obtain the context-aware representations of each word:

$$h_t = BiLSTM(w_t, h_{t-1}) \tag{12}$$

Apparently, we can get the sequence of the context-aware representations $np_t = (h_i, h_2, \ldots, h_j)$ from the outputs of the BiLSTM, where h_i means the start of the candidate antecedent, and h_j means the end of the candidate antecedent. we use h_i and h_j as the context-aware boundary representations of the candidate antecedents.

Then, we use a simple and effective scaled dot-product-based attention layer to calculate the weight distribution of the words in the candidate antecedent. We regard the zero pronoun representations h_{zp} as the query term, and regard the context-aware representations of all words in the candidate antecedent as the key term and value term. For simplicity in formula expression, we use the matrix H_{np} to denote he key term and value term:

$$h_{np}^{(attn)} = Softmax(\frac{h_{zp} H_{np}^T}{\sqrt{d_{np}}}) H_{np} \tag{13}$$

where $h_{zp} \in \mathbb{R}^{d_{zp}}$, $H_{np} \in \mathbb{R}^{n \times d_{np}}$, $d_{zp} = d_{np}$, n denotes the number of words in the candidate antecedent. d_{np} denotes the dimension of context-aware representations of all words in the candidate antecedent. d_{zp} denotes the dimension of zero pronoun representations. $h_{np}^{(attn)}$ is the weighted sum the context-aware representations of all words in the candidate antecedent, where $h_{np}^{(attn)} \in \mathbb{R}^{d_{np}}$.

Finally, we take the concatenate of h_i, h_j, and $h_{np}^{(attn)}$ as the the final representations of each candidate antecedent.

$$h_{np} = [h_i; h_j; h_{np}^{(attn)}] \tag{14}$$

3.4 Discrimination Module

After obtaining the representations of the zero pronoun and all candidate antecedents of this zero pronoun, we can feed these representations into the discrimination module to predict the real referent of the current zero pronoun.

For the discrimination module, this paper uses a bilinear function to calculate the probability distribution of all candidate antecedents of the current zero pronoun.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 832-842, Hainan, China, October 31 - November 1, 2020.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

$$P(np_t|w_{zp}) = Softmax(h_{zp}UM_{np}^T + b) \tag{15}$$

$$\sum_{t=1}^{m} P(np_t|w_{zp}) = 1 \tag{16}$$

The parameters of the bilinear function are U and b, where $U \in \mathbb{R}^{k\times k}$, $b \in \mathbb{R}^k$, k denotes the dimension of the input vector of the bilinear function. h_{zp} denotes the zero pronoun representations. M_{np} denotes the matrix of all candidate antecedents of the current zero pronoun, where $h_{zp} \in \mathbb{R}^{1\times k}$, $M_{np} \in \mathbb{R}^{m\times k}$. m denotes the number of all candidate antecedents of the current zero pronoun.

Given the probability distribution of of all candidate antecedents of the current zero pronoun, we select the candidate antecedent with the highest probability as the real referent of the current zero pronoun.

3.5 The Integration of Semantic Representations

To make better use of the semantic representations from the pre-trained semantic dependency graph parser, We integrate the semantic representations of word-level and sentence-level into the Chinese zero pronoun resolution model.

Inspired by the work of Jiang et al. (Jiang et al., 2020), we first extract all output vectors from the BiLSTM-based encoder of the pre-trained semantic dependency graph parser and then use a set of trainable parameters to weighted sum these vectors to obtain the final semantic representations. we use h_t^{sem} to denote the semantic representations of a word. This process is formally denoted by the following formula:

$$h_t^l = BiLSTM^{(l)}(w_t, h_{t-1}) \tag{17}$$

$$h_t^{(sem)} = \sum_{l=1}^{L} \alpha_l h_t^l \tag{18}$$

where w_t is the original word representations，L is the layer number of the Bi-LSTM-based encoder, and α_l is the normalized weight of each layer.

For the integration of the word-level semantic representations, we simply concatenate the semantic representations of each word with its original word representations:

$$w_t^{(sem)} = [w_t; h_t^{(sem)}] \tag{19}$$

For the integration of the sentence-level semantic representations, We use the CNN-based sentence-level semantic representations extractor to perform 2-dimensional convolution and hierarchical pooling operations on the sentence sequence. Hierarchical pooling (Shen et al., 2018) is a combination of average pooling and max-pooling, which has better ability to capture word-order information. we use S_1^n to denote a sentence sequence with n words. This process is shown in the following formulas:

$$s^{(sem)} = Pooling(Convolution(S_1^n)) \tag{20}$$

After we obtain the sentence-level semantic representations, we integrate it into the zero pronoun module and the candidate antecedent module. We use two different multi-layer perceptrons to transform sentence-level semantic representations into zero pronoun-related and candidate antecedent-related representations. In this way, even if the zero pronoun and candidate antecedent are in the same sentence, these sentence-level semantic representations are different. This process is shown in the following formulas:

$$h_{zp}^{(sem)} = MLP^{(zp)}(s^{(sem)}) \tag{21}$$

$$h_{np}^{(sem)} = MLP^{(np)}(s^{(sem)}) \tag{22}$$

Finally, the zero pronoun representations and candidate antecedent representations that are integrated into the semantic representations can be formalized as:

$$h_{zp} = [h^{(pre)}; h^{(fol)}; h^{(sem)}] \tag{23}$$

$$h_{np} = [h_i; h_j; h_{np}^{(attn)}; h^{(sem)}] \tag{24}$$

3.6 Training Objective

The training objective is defined as:

$$Loss = -\Sigma_{zp} log P(np_t|w_{zp}) \tag{25}$$

where zp means the number of all anaphoric zero pronouns in the training set.

4 Experiment

4.1 Dataset and Resource

We conduct our experiments on the OntoNotes-5.0 dataset[0] which consists of document-level text selected from 6 domains : Broadcast News(BN), Newswire(NW), Broadcast Conversation(BC), Web Blog (WB), Telephone Conversation (TC) and Magazine(MZ). The training set has 1391 documents, a total of 36487 sentences and 12111 zero pronouns; The development set has 172 documents with a total of 6083 sentences and 1713 zero pronouns. The pre-trained word embedding used in Chinese zero pronoun resolution are trained by Word2Vec algorithm on Chinese Gigawords [1]. For Pre-training the Chinese semantic dependency graph parser, we use the SemEval-2016 Task 9 dataset[2]. For BERT related experiments, We use the Chinese Bert-base model, which has been pre-trained by the Google [3].

4.2 Evaluation Measures

We adopt the Recall, Precision and F-score (denoted as F) as the evaluation metrics of our Chinese zero pronoun resolution model. More specifically, recall, precision and F are defined as:

$$P = \frac{the\ number\ of\ zero\ pronouns\ predicted\ correctly}{the\ number\ of\ all\ predicted\ zero\ pronouns} \tag{26}$$

$$R = \frac{the\ number\ of\ zero\ pronouns\ predicted\ correctly}{the\ number\ of\ zero\ pronouns\ labeled\ in\ all\ datasets} \tag{27}$$

$$F = \frac{2PR}{P+R} \tag{28}$$

4.3 Hyperparameters

For Zero Pronoun Module, the hidden dimension of the LSTM is 128 and the output dimension of the affine-based attention layer is 128. For Candidate Antecedents Module, the hidden dimension of the LSTM is 128 and the output dimension of the scaled dot-product-based attention layer is 128. For all pre-trained representations, we convert the final input dimension to 256. For all LSTM, dropout rates are set to 0.33. For other neural network, dropout rates are set to 0.5. For training, the model is optimized by the Adam algorithm with the initial learning rate 0.003.

[0]http://catalog.ldc.upenn.edu/LDC2013T19
[1]https://catalog.ldc.upenn.edu/LDC2003T09
[2]https://github.com/HIT-SCIR/SemEval-2016
[3]https://github.com/google-research/bert

4.4 Main experiments

We chose three deep learning-based Chinese zero pronoun resolution model implemented by Yin et al as reference: Deep Memory Network-based Chinese zero pronoun resolution model (Yin et al., 2017)(DMN-ZP Model), Self-attention-based Chinese zero pronoun resolution model (Yin et al., 2018b)(SA-ZP Model) and Deep Reinforcement Learning-based Chinese zero pronoun resolution model (Yin et al., 2018a)(DRL-ZP Model).

We evaluate the performance of our Chinese zero pronoun resolution model on OntoNotes-5.0 development dataset with two different model settings: Chinese zero pronoun resolution model without pre-trained Chinese semantic dependency graph parser(Our Baseline Model), Chinese zero pronoun resolution model with pre-trained Chinese semantic dependency graph parser(Our Semantic-aware Model). The specific experimental results are shown in Table 1:

Model	NW(84)	MZ(162)	WB(284)	BN(390)	BC(510)	TC(283)	Overall
DMN-ZP Model	48.8	46.3	59.8	58.4	53.2	**54.8**	54.9
DRL-ZP Model	63.1	50.2	63.1	56.7	57.5	54	57.2
SA-ZP Model	**64.3**	52.5	62	**58.5**	57.6	53.2	57.3
Our Baseline Model	63.3	51.5	61.8	58.2	57.5	53.1	57.2
Our Semantic-aware Model	64.3	**52.7**	**63.3**	58.3	**58.8**	53.1	**57.6**

Table 1: Comparison of Different Chinese Zero pronoun Resolution Models

Compared with the baseline model, our semantic-aware model has achieved a 0.4 % improvement in F-score. Compared with previous deep learning-based models, the performance of our semantic-aware model is the best. According to the experimental results in various fields, we found that our semantic-aware model obtains the highest F-score in the MZ, BC and WB fields. Among them, the improvement of our semantic-aware model in the BC field is the most obvious. However, in the field of NW, BN and TC, the performance of our semantic-aware model has no advantage. One possible reason for this phenomenon is that the performance of the semantic dependency graph parser in these three fields is relatively poor, and it cannot provide valuable semantic information to the task of Chinese zero pronoun resolution.

4.5 Ablation Experiment

In order to further verify the effectiveness of our model, we tested the performance of models using the word-level and sentence-level integration method through ablation experiments. According to the experimental results in Table 2, we found that both integration methods can improve the performance of our model, and when both integration methods are used simultaneously, the performance of our model is optimal. The word-level integration method can only focus on the semantic information within the same sentence, while the sentence-level integration method has the ability to focus on the difference in sentence-level semantic information between different sentences. Therefore, the word-level integration method may be more suitable for the case where the zero pronoun and the candidate antecedent are in the same sentence, and the sentence-level integration method is more suitable for the case where the zero pronoun and the candidate antecedent are in different sentences. It is the complementarity of these two methods that makes the performance of our model continuously improved.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 832-842, Hainan, China, October 31 - Novermber 1, 2020.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	Overall
Baseline Model	57.2
Sematic-aware Model(Sentence-Level)	57.3
Sematic-aware Model(Word-Level)	57.5
Sematic-aware Model	**57.6**

Table 2: Ablation experiment results

4.6 Integration with BERT

BERT (Devlin et al., 2018) is a pre-trained language model with strong capabilities and wide application. Many BERT-based natural language processing models have achieved the state of the art performance. In order to verify the effectiveness of our model after integrating the BERT representations, we compared and analyzed the following four sets of experiments: Baseline Model without BERT, Baseline model with BERT, Semantic-aware Model without BERT, Semantic-aware Model with BERT. It is worth noting that the method of integrating BERT information is the same as the method of integrating semantic dependency graph information. The specific experimental results are shown in Table 3:

Model	Overall
Baseline Model without BERT	57.2
Baseline Model with BERT	57.7
Semantic-aware Model without BERT	57.6
Semantic-aware Model with BERT	**57.9**

Table 3: Integration with BERT

According to the experimental results in the Table 3, we can see that the performance of the Semantic-aware Model with BERT is the best. This shows that BERT information and semantic dependency graph information have certain complementarity in the Chinese zero pronoun resolution task. But by comparing the performance of the Semantic-aware Model without BERT and Baseline model with BERT, We can see that the BERT information contributes more to the Chinese zero pronoun resolution task than the semantic dependency graph information. In addition, we can also see that BERT information improves the Baseline Model more than the Semantic-aware Model. This shows that the BERT model may encode part of the semantic information of the semantic dependency graph. Based on the above analysis, we hope that in the future research, we can further integrate the semantic dependency graph and even the information of semantic role labeling on the basis of the BERT model, so as to further enhance the ability of the BERT model in the Chinese zero pronoun resolution task.

4.7 Conclusion

This paper proposes a semantic-aware Chinese zero pronoun resolution model with pre-trained semantic Dependency Parser. In order to effectively integrate semantic information from the pre-trained semantic dependency graph parser, We integrate semantic representations into the Chinese zero pronoun resolution model at two levels: word level and sentence level. The experimental results show that our proposed model achieves better performance than other deep learning-based models. In addition, we find

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

that BERT information and semantic dependency graph information have certain complementarity in the Chinese zero pronoun resolution task. After our model is enhanced with the BERT representations, its performance has been further improved. In future research, we will explore the integration of BERT information and semantic dependency graph information to provide richer information for Chinese zero-finger resolution tasks.

Acknowledgements

This research project is supported by the National Natural Science Foundation of China (61872402), the Humanities and Social Science Project of the Ministry of Education (17YJAZH068)，Science Foundation of Beijing Language and Culture University (supported by the Fundamental Research Funds for the Central Universities) (18ZDJ03)，the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (UCCA). page 11.

Wanxiang Che, Yanqiu Shao, Ting Liu, and Yu Ding. 2016. SemEval-2016 task 9: Chinese semantic dependency parsing. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1074–1080. Association for Computational Linguistics.

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. 1:778–788.

Jacob Devlin, Mingwei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv: Computation and Language*.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia, July. Association for Computational Linguistics.

Yuze Gao, Yue Zhang, and Tong Xiao. 2017. Implicit syntactic features for targeted sentiment analysis. page 9.

Wei Jiang, Zhenghua Li, and Min Zhang. 2020. Syntax-enhanced ucca semantic parsing. *Beijing Da Xue Xue Bao*, 56(1):89–96.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for chinese zero anaphora resolution. pages 882–891.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution.

Ting Liu, Yiming Cui, Qingyu Yin, Weinan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. 1:102–111.

Converse S P and Palmer M S. 2006. *Pronominal anaphora resolution in Chinese*. University of Pennsylvania.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms.

Zizhuo Shen, Huayong Li, Dianqing Liu, and Yanqiu Shao. 2019. Dependency-gated cascade biaffine network for chinese semantic dependency graph parsing. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 840–851. Springer.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. Chinese zero pronoun resolution with deep memory network. pages 1309–1318.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018a. Deep reinforcement learning for chinese zero pronoun resolution. 1:569–578.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018b. Zero pronoun resolution with attention-based neural network. pages 13–23.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural rst parsing with implicit syntax features. pages 559–570.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2017. End-to-End Neural Relation Extraction with Global Optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1730–1740, Copenhagen, Denmark. Association for Computational Linguistics.

Meishan Zhang, Zhenghua Li, Guohong Fu, and Min Zhang. 2019. Syntax-enhanced neural machine translation with syntax-aware word representations. pages 1151–1161.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of chinese zero pronouns: A machine learning approach. pages 541–550.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China