Hierarchical Attention Network with Pairwise Loss for Chinese Zero Pronoun Resolution

Peiqin Lin,[1] **Meng Yang**[1,2]*
[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[2]Key Laboratory of Machine Intelligence and Advanced Computing (SYSU), Ministry of Education
linpq3@mail2.sysu.edu.cn, yangm6@mail.sysu.edu.cn

Abstract

Recent neural network methods for Chinese zero pronoun resolution didn't take bidirectional attention between zero pronouns and candidate antecedents into consideration, and simply treated the task as a classification task, ignoring the relationship between different candidates of a zero pronoun. To solve these problems, we propose a Hierarchical Attention Network with Pairwise Loss (HAN-PL), for Chinese zero pronoun resolution. In the proposed HAN-PL, we design a two-layer attention model to generate more powerful representations for zero pronouns and candidate antecedents. Furthermore, we propose a novel pairwise loss by introducing the correct-antecedent similarity constraint and the pairwise-margin loss, making the learned model more discriminative. Extensive experiments have been conducted on OntoNotes 5.0 dataset, and our model achieves state-of-the-art performance in the task of Chinese zero pronoun resolution.

Introduction

Zero pronoun, as a special linguistic phenomenon in pro-dropped languages, is pervasive in Chinese documents (Zhao and Ng 2007). A zero pronoun is a gap in the sentence, which refers to the component that is omitted because of the coherence of language. As Figure 1 is shown, a zero pronoun can be an anaphoric zero pronoun (AZP) if it corefers to one or more mentions in the associated text, which is usually represented by a coreference chain, or a non-anaphoric one if there are no such mentions. In this example, $*pro*_1$ is anaphoric and corefers to the mention "The police", while $*pro*_2$ is non-anaphoric. These mentions for interpreting zero pronouns are called the antecedents. How to correctly resolve AZP is a challenging topic in semantic understanding and has attracted much attention.

Early approaches employed rule-based methods to resolve Chinese zero pronoun resolution (Converse and Palmer 2006; Yeh and Chen 2007). After that, some traditional machine learning models with hand-crafted features, including supervised approaches and unsupervised approaches, were extensively employed to solve the problem

[警方] 怀疑 这是 一起 黑枪 案件，***pro*$_1$** 将 枪械 和 皮包 交送 市里 ***pro*$_2$** 以 清理 案情。
[The police] suspected that this is a criminal case about illegal guns, ***pro*$_1$** brought the guns and bags to the city ***pro*$_2$** to deal with the case.

Figure 1: An example of zero pronoun phenomenon. Zero pronouns are denoted as "*pro*".

(Zhao and Ng 2007; Kong and Zhou 2010; Chen and Ng 2013; 2014; 2015). (Zhao and Ng 2007) investigated a series of syntactic features based on parse trees to locate and resolve zero anaphoras. Based on (Zhao and Ng 2007), (Chen and Ng 2013) further introduced lexical features and coreference links between zero pronouns. Despite the effectiveness of feature engineering, it is labor intensive and highly relies on annotated corpus.

Due to the powerful ability of deep learning, (Chen and Ng 2016) was the first to apply a deep neural network for the task. Then (Liu et al. 2016) produced pseudo dataset and adopted a pre-training-then-adaptation method; (Yin et al. 2017) introduced a memory-based network to choose the correct antecedent for the specific zero pronoun. To capture more information, (Yin et al. 2016) encoded both local information and global information for candidates; (Yin et al. 2018a) integrated local and global decision-making by exploiting deep reinforcement learning models. In addition, self-attention mechanism has also been introduced for encoding zero pronouns and the attention-based recurrent neural network was applied for encoding candidate antecedents by their content (Yin et al. 2018b). However, these methods either did not consider any interaction between zero pronouns and candidate antecedents (Chen and Ng 2016; Yin et al. 2018a) or just employed unidirectional attention from the representations of zero pronouns to those of candidate antecedents (Liu et al. 2016; Yin et al. 2018b), weakening the representation ability of the learned features. Moreover, these methods simply formulate the resolution task as a classification task (e.g., whether a candidate is the an-

*Corresponding Author

tecedent of a zero pronoun), which ignores the relationship between different candidates of a zero pronoun (e.g., the correct candidates are similar and their scores should be larger than those of wrong candidates by a large margin).

To address these issues, we propose a novel framework, Hierarchical Attention Network with Pairwise Loss (HAN-PL). We design a two-layer hierarchical attention network, which not only takes bidirectional attention into consideration to solve the task firstly, to generate more powerful representations, but also presents a pairwise loss that integrates more discrimination into the learned model. Hierarchical Attention Network (HAN) employs interactive attention and self attention to better model zero pronouns and candidate antecedents, while Pairwise Loss (PL) integrates correct-antecedent similarity into pairwise-margin loss. The experiments on OntoNotes 5.0 clearly show that the proposed HAN-PL outperforms all of the baseline systems and gains state-of-the-art performance significantly. The major contributions of this paper are three-fold:

- Learning more powerful representations of zero pronouns and candidate antecedents interactively with the proposed Hierarchical Attention Mechanism;
- Guiding the optimization of the model with a pairwise-margin loss, which is more reasonable than cross entropy loss used in previous methods;
- Taking the constraint of correct-antecedent similarity into account for utilizing the global information provided by the chain information.

Related Work

In this section, we firstly give a brief summary of early efforts for attention mechanism and max-margin loss, which are related to our contributions, and then we briefly review the popular approaches for Chinese zero pronoun resolution.

Attention Mechanism for Natural Language Processing

(Bahdanau, Cho, and Bengio 2014) was the first to apply attention mechanism in natural language processing (NLP). Since then, attention mechanism has been widely used in many NLP tasks, such as document classification (Yang et al. 2016), machine reading comprehension (Kadlec et al. 2016) and so on. Some approaches for reading comprehension (Seo et al. 2016; Wang, Yan, and Wu 2018), which proposed various interaction ways between question and passage, really inspired us.

Max-Margin Loss

Max-margin loss is more reasonable than cross entropy loss in some tasks, like image similarity (Wang et al. 2014) and face recognition (Schroff, Kalenichenko, and Philbin 2015). Actually, max-margin loss (Wiseman et al. 2015; Clark and Manning 2016) has also been adopted to coreference resolution, a similar task to the task in this paper. However, the designed loss for coreference resolution requires careful tuning and isn't suitable for model optimization.

Zero Pronoun Resolution for Chinese

Previous approaches for Chinese zero pronoun resolution modeled the task with traditional machine learning methods or deep learning methods, and then trained the model with cross entropy loss.

Recently, some deep learning models has been applied for Chinese zero pronoun resolution (Chen and Ng 2016; Yin et al. 2017). (Yin et al. 2018b) introduced the self-attention mechanism for encoding zero pronouns and the attention-based recurrent neural network for encoding candidate antecedents by their contents, respectively. Moreover, (Yin et al. 2018b) treated the resolution task as a classification task and guided the optimization with cross entropy loss:

$$l_{ce} = -\delta(zp, np) * log(g(zp, np)) \quad (1)$$

where $g(zp, np) \in [0, 1]$, which is computed by (Yin et al. 2018b), is the coreference probability of the given zero pronoun zp and its candidate antecedent np. $\delta(zp, np)$ represents the actual coreference result between zp and np: if they are coreference, $\delta(zp, np) = 1$ or otherwise, $\delta(zp, np) = 0$.

However, the above method ignores the information of candidate antecedents when encoding zero pronouns, which weaken the representation ability of the learned features of zero pronouns and candidate antecedents. In addition, the cross entropy loss used in the method can't ensure that the resolution scores of correct candidates are larger than those of wrong candidates by an enough margin.

Model

To implement the task of Chinese zero pronoun resolution in a more reasonable way, we propose a Hierarchical Attention Network with Pairwise Loss (HAN-PL). In the proposed HAN-PL, we design a two-layer attention model to generate more powerful representations for zero pronouns and candidate antecedents. In addition, we integrate the constraint of similarities among correct antecedents into pairwise-margin loss, for guiding the training of the model. In this section, we firstly give the description of the resolution task, and then describe our major contributions, namely Hierarchical Attention Network and Pairwise Loss, in detail.

Task Description

In the problem of Chinese zero pronoun resolution, the positions of zero pronouns have been already given by the former step of zero pronoun detection (Kong and Ng 2013). Given an anaphoric zero pronoun zp, candidate antecedents $S_{zp} = \{np_1, np_2, ..., np_k\}$ are extracted by capturing maximal or modifier noun phrases that are at most two sentences away from zp (Chen and Ng 2015), which recalls most (about 98%) of the antecedents. In addition, the context $\{npc_1, npc_2, ..., npc_k\}$ for antecedents are also considered. To determine the correct antecedent of zp, a Hierarchical Attention Network $f(zp, np)$ is designed, and will be detailed in the following subsection.

Hierarchical Attention Network

Previous methods don't take enough information into account, e.g., they don't consider enough interaction between

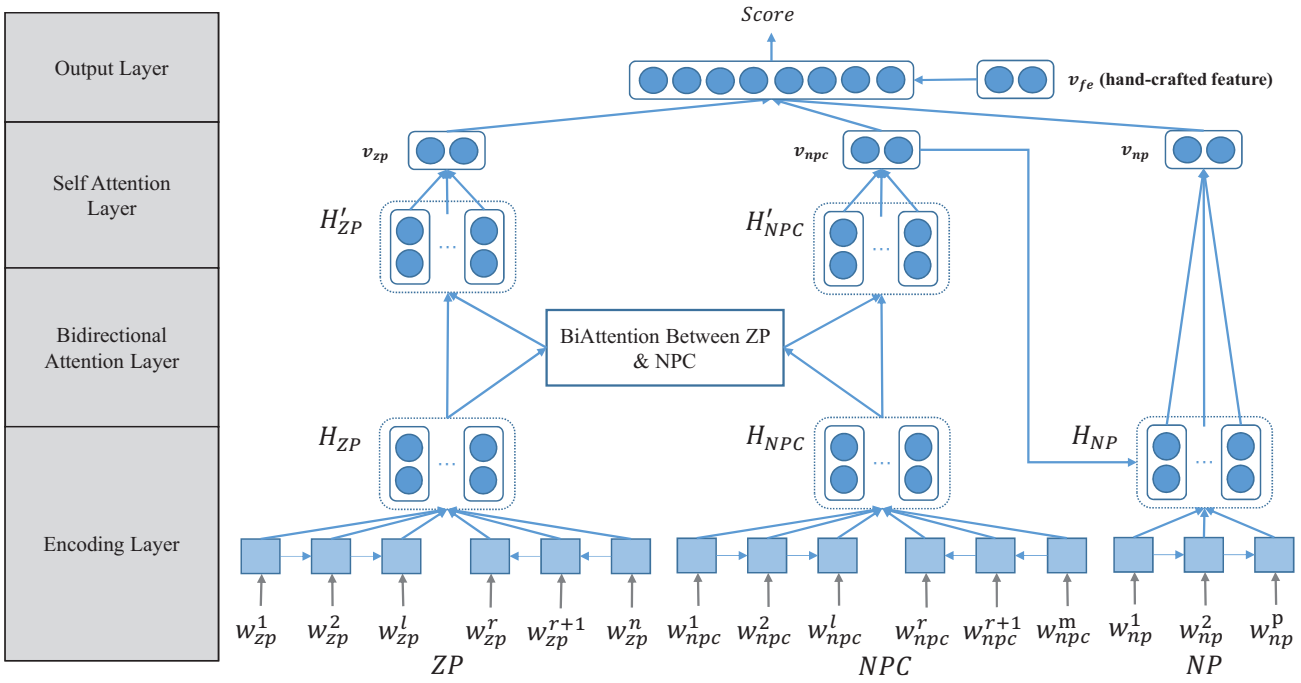

Figure 2: The overall architecture of HAN. $\boldsymbol{v}_{zp}$, $\boldsymbol{v}_{npc}$ and $\boldsymbol{v}_{np}$ are generated by our attention model, and $\boldsymbol{v}_{fe}$ is a hand-crafted feature (Chen and Ng 2013; 2016), which is used in previous methods (Yin et al. 2018a; 2018b) and can improve the performance.

zero pronouns and candidate antecedents. The contexts of antecedents, which may be helpful for modeling, is usually ignored in previous methods. In addition, the information of antecedents is rarely considered to assist the modeling of zp. To model zp, np and npc in a better way, we design a Hierarchical Attention Network, as shown in Figure 2.

Modeling Context of Zero Pronoun and Candidate Antecedent As zero pronouns are gaps that have no text, the context of zp will be taken as inputs to model zp (Yin et al. 2017; 2018a; 2018b). In addition, we utilize the information of the context of candidate antecedent npc, which is ignored in previous method and should be considered actually. Here we apply a two-layer attention network, including a bidirectional attention layer and a self attention layer, to learn the representation of zp and npc interactively.

Encoding Layer encodes zp and npc with vanilla recurrent neural networks (RNNs) which is applied in previous method (Yin et al. 2018b). We firstly use a left-to-right RNN for encoding the left context of zp, and a right-to-left RNN for encoding the right context. After encoding, we can get the hidden states of the preceding and following context for zp, respectively. Therefore, we can get the final hidden states $\boldsymbol{H}_{zp} \in R^{d \times n}$ of zp by simply concatenating the two matrices, where d is the hidden size and n is the length of the context of zp. In addition, we also use the same method to encode the context of candidate antecedent npc. The final states of npc are denoted as $\boldsymbol{H}_{npc} \in R^{d \times m}$, where m is the length of the context of antecedents.

Bidirectional Attention Layer learns the representations of zp and npc in an interactive way. The layer calculates an attention matrix firstly:

$$\boldsymbol{Att} = ReLU(\boldsymbol{W}_l^T \boldsymbol{H}_{zp})^T \cdot ReLU(\boldsymbol{W}_l^T \boldsymbol{H}_{npc}) \quad (2)$$

where $\boldsymbol{W}_l^T$ is a trainable matrix for linear transformation, $\boldsymbol{Att} \in R^{n \times m}$ and $\boldsymbol{Att}(i,j)$ represents the attention score between the ith word of zp and the jth word of npc.

With the calculated attention matrix $\boldsymbol{Att}$, we can get the normalized scores from npc to zp, signifying which words of npc are most relevant to each word of zp:

$$\alpha_{ij} = \frac{exp(\boldsymbol{Att}(i,j))}{\sum_{k=1}^{n} exp(\boldsymbol{Att}(k,j))} \quad (3)$$

The aligned representation from npc to the ith word of zp can thus be derived as:

$$\tilde{\boldsymbol{H}}_{zp}(i) = \sum_{j=1}^{m} \alpha_{ij} * \boldsymbol{H}_{np}(j) \quad (4)$$

Finally, we combine the original contextual representations and the corresponding attention vectors of zp by simply summing up them:

$$\boldsymbol{H}'_{zp} = \boldsymbol{H}_{zp} + \tilde{\boldsymbol{H}}_{zp} \quad (5)$$

Similar to calculate the final representations of zp, we can also get the representations of npc, namely, $\boldsymbol{H}'_{npc}$.

Self Attention Layer is finally applied on the representations of zp and npc to get the final vectors, respectively. We can calculate attention scores for zp as follows:

$$\boldsymbol{Sco} = softmax(\boldsymbol{W}_1 tanh(\boldsymbol{W}_2 \boldsymbol{H}'_{zp})) \quad (6)$$

where $\boldsymbol{W}_1 \in R^{1\times d}$ and $\boldsymbol{W}_2 \in R^{d\times d}$ are the weight matrices. And then we can get the final vector of zp:

$$\boldsymbol{v}_{zp} = \boldsymbol{Sco} \cdot (\boldsymbol{H}'_{zp})^T \tag{7}$$

Similarly, we can get the final representation of npc, namely $\boldsymbol{v}_{npc}$.

Modeling Content of Candidate Antecedent With no doubt, the content of candidate antecedents should be also considered. Similar to the encoding of context, we apply a vanilla recurrent neural network, whose input is comprised by the words in the candidate antecedent (Yin et al. 2018b). Then we can get the hidden states $\boldsymbol{H}_{np} = \{\boldsymbol{h}^1_{np}, \ldots, \boldsymbol{h}^i_{np}, \ldots, \boldsymbol{h}^p_{np}\}$ for noun pronoun content np, where p is the length of np.

To better capture the more informative parts of the content of candidate antecedents, we here integrate an attention layer into our model by utilizing the information of its context:

$$\beta_i = softmax(\boldsymbol{W}_{att}[\boldsymbol{h}^i_{np}; \boldsymbol{v}_{npc}] + \boldsymbol{b}_{att}) \tag{8}$$

where $\boldsymbol{W}_{att}$ and $\boldsymbol{b}_{att}$ are weight matrix and bias, and then we can get the final representation $\boldsymbol{v}_{np}$:

$$\boldsymbol{v}_{np} = \sum_{i=1}^{p} \beta_i \boldsymbol{h}^i_{np} \tag{9}$$

Getting Resolution Results After generating the representations of zp, np and npc, we calculate the resolution score for each zero pronoun-candidate antecedent by using a two-layers feed-forward neural network. Taking $\boldsymbol{v}_{zp}$, $\boldsymbol{v}_{npc}$ and $\boldsymbol{v}_{np}$ as inputs, our model calculate the resolution score by going through two $tanh$ layers:

$$\boldsymbol{r}_j = tanh(\boldsymbol{W}_j \boldsymbol{r}_{j-1} + \boldsymbol{b}_j) \tag{10}$$

where $\boldsymbol{W}_j$ and $\boldsymbol{b}_j$ are the parameters of this feed-forward neural network, $\boldsymbol{r}_0 = (\boldsymbol{v}_{zp}; \boldsymbol{v}_{np}; \boldsymbol{v}_{npc}; \boldsymbol{v}_{fe})$. The handcrafted feature $\boldsymbol{v}_{fe}$, which is used in previous work (Yin et al. 2016; 2017; 2018a; 2018b), is designed to capture the syntactic, position and other relations between zp and np (Chen and Ng 2013; 2016). Then we can get the resolution score:

$$s_i = \boldsymbol{W}_s \boldsymbol{r}_{-1} + \boldsymbol{b}_s \tag{11}$$

where $s_i \in (-\infty, \infty)$ is a scalar which denotes the resolution probability of the ith candidate np_i being predicted to be the antecedent, and $\boldsymbol{r}_{-1}$ is the output of the second hidden layer. After that, we obtain the resolution scores for all the candidates $\{s_1, s_2, ..., s_k\}$. The candidate with the biggest score is selected to be the antecedent of zp.

Pairwise Loss

To guide the optimization of the model, we design a reasonable loss named Pairwise Loss, which is based on a pairwise-margin loss and a similarity constraint, instead of cross entropy loss used in previous methods. We call our loss function Pairwise Loss for two major reasons:

- We take each correct antecedent and each wrong antecedent in the candidate set as a pair, and then compute the pairwise-margin loss between them;
- We take correct antecedents in pair, and then design a similarity constraint for better training the model.

Pairwise-Margin Loss Previous methods treated the task as a coreference classification task of each zero pronoun-candidate antecedent pair, namely classifying the examples into two categories, coreference or not coreference, and then trained their model by minimizing the cross entropy error, which is less reasonable. Firstly, the cross entropy loss function set a fixed decision boundary for all the examples, which is not flexible enough. Secondly, the examples which is not coreference are much more than the examples which is coreference in the task of coreference resolution, and it will lead to the problem of imbalanced data, which can't be solved effectively in a classification task.

To solve the above issues, we design a pairwise-margin loss, which is more reasonable. In the extracted candidate set $S_{zp} = \{np_1, np_2, ..., np_k\}$ for zero pronoun zp, we can simply divide it to two sets, the correct candidate set $S^T_{zp} = \{np_1, np_2, ..., np_{k_1}\}$ and the wrong candidate set $S^F_{zp} = \{np_1, np_2, ..., np_{k_2}\}$. Then we can design the original loss function for different cases (shown in Eq.(12) and illustrated in Figure 3) as follows:

- **Case 1:** If the candidate set contains both correct antecedents and wrong antecedents ($S^T_{zp} \neq \emptyset \wedge S^F_{zp} \neq \emptyset$), we design a pairwise-margin loss, where m is the margin between correct antecedents and wrong antecedents;
- **Case 2:** If the candidate set only contains correct antecedents ($S^T_{zp} \neq \emptyset \wedge S^F_{zp} = \emptyset$, named **Case 2a**) or wrong antecedents ($S^T_{zp} = \emptyset \wedge S^F_{zp} \neq \emptyset$, named **Case 2b**), boundary values, namely the lower value bv_T for the former case and the upper value bv_F for the latter, will be set to guide the training for these examples. The boundary values are set according to the resolution scores generated by the samples which satisfy **Case 1**[1];
- **Case 3:** If the candidate set is empty ($S^T_{zp} = \emptyset \wedge S^F_{zp} = \emptyset$), the corresponding zp will be ignored.

With the designed pairwise-margin loss, the two issues we mentioned above can be solved well. Instead of setting a clear decision boundary, the pairwise-margin loss requires that the resolution probabilities of the correct antecedents are somewhat higher than those of the wrong antecedents. In addition, pairwise-margin loss can solve the problem of imbalanced data easily. For the first case, namely the most common case, the number of correct examples used is the same as the number of wrong examples while computing pairwise-margin loss.

Correct-Antecedent Similarity Since the correct antecedents in the candidate set of the specific zp must have same or close meanings, we integrate the similarities among correct antecedents into the pairwise-margin loss function mentioned above. We compute the cosine similarities among correct antecedents, and then define the constraint:

$$L_c = \sum_{zp\in ZP} \sum_{x_1 \in S^T_{zp}} \sum_{x_2 \in S^T_{zp}} (1 - sim(\boldsymbol{v}_{x_1}, \boldsymbol{v}_{x_2})) \tag{13}$$

[1]We firstly used part of the data which can be applied with pairwise-margin loss directly to conduct preliminary training, and then obtained the lower bound of the correct examples and the upper bound of the wrong examples.

$$L_o = \sum_{zp \in ZP} \begin{cases} \sum_{x_1 \in S^T_{zp}} \sum_{x_2 \in S^F_{zp}} |f(zp, x_2) - f(zp, x_1) + m|_+ & \text{if } S^T_{zp} \neq \emptyset \wedge S^F_{zp} \neq \emptyset \\ \sum_{x \in S^T_{zp}} |bv_T - f(zp, x)|_+ & \text{if } S^T_{zp} \neq \emptyset \wedge S^F_{zp} = \emptyset \\ \sum_{x \in S^F_{zp}} |f(zp, x) - bv_F|_+ & \text{if } S^T_{zp} = \emptyset \wedge S^F_{zp} \neq \emptyset \\ 0 & \text{if } S^T_{zp} = \emptyset \wedge S^F_{zp} = \emptyset \end{cases} \tag{12}$$

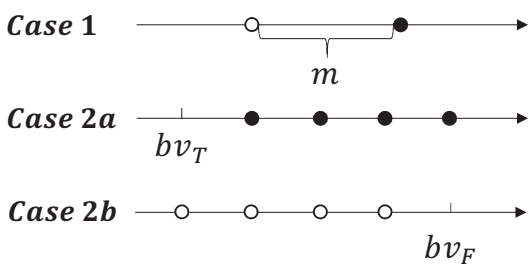

Figure 3: Illustration of the pairwise-margin loss. The black dots and white dots denote the scores of the correct and wrong candidate antecedents, respectively.

	N_d	N_s	N_w	N_{azp}
Train	1,391	36,487	756K	12,111
Test	172	6,083	110K	1,713

Table 1: Statistics on the train and test dataset. N_d, N_s, N_w and N_{azp} denote the numbers of documents, sentences, words and anaphoric zero pronouns, respectively.

where $sim(\cdot)$ is the function which computes cosine similarity between $\boldsymbol{v}_{x_1}$ and $\boldsymbol{v}_{x_2}$, which are the representations of np described in Section .

Final Loss Function The model is trained by minimizing the combination of the pairwise-margin loss, the similarity constraint and a L_2 regularization term:

$$L = L_o + \lambda_c L_c + \lambda \|\theta\|_2^2 \tag{14}$$

where λ_c and λ are the weight of L_c and L_2 regularization term.

In the designed L, the pairwise margin loss L_o is designed to make all correct antecedents get bigger resolution scores than wrong antecedents, similarity constraint L_c makes correct antecedents of the same zero pronoun have similar representations, and $\|\theta\|_2^2$ is L_2 regularization term for avoiding over-fitting. With the proposed loss L, the resolution task can be solved in a reasonable way.

Experiment

Experiment Setup

Datasets We conduct experiments on the Chinese part of the OntoNotes 5.0 dataset. Documents in this dataset are from 6 sources: **BN** (Broadcast News), **NW** (Newswire), **BC** (Broadcast Conversation), **WB** (Web Blog), **TC** (Telephone Conversation) and **MZ** (Magazine). The statistics of dataset are reported in Table 1.

Metrics Following previous methods on zero pronoun resolution (Zhao and Ng 2007; Chen and Ng 2016; Yin et al. 2016; 2017; 2018a; 2018b), F-score (F) is employed to evaluate our model and it is calculated as follows:

$$F = \frac{2PR}{P+R},\ P = \frac{N_{hit}}{N_{azp^*}},\ R = \frac{N_{hit}}{N_{azp}} \tag{15}$$

where P and R are the precision and recall of the model, N_{hit}, N_{azp^*} and N_{azp} denote the numbers of the examples which are correctly resolved, the examples which have non-empty candidate sets and AZPs in the test set, respectively. We report the F-scores for each source in addition to the overall result.

Baselines We use the recent zero pronoun resolution methods for Chinese as our baselines, namely, a learning-based model (Zhao and Ng 2007); an unsupervised method (Chen and Ng 2015); and others are deep-learning-based methods (Chen and Ng 2016; Yin et al. 2016; 2017; Liu et al. 2016; Yin et al. 2018a; 2018b).

Hyperparameters We minimize the loss-function by Adam (Kingma and Ba 2014) with learning rate 5e-5 and L_2 weight 1e-4. The input embedding vector dimension is 100, the dimension d of hidden layers and the representations is 256, the margin m of pairwise-margin loss is 0.1, the boundary values for correct antecedents and wrong antecedents are 0.3 and 0.4, and the weight of the similarity constraint λ_c is 0.5. Besides, we add the dropout (Hinton et al. 2012) with a probability of 50% on the output of each layer.

Same to previous methods (Yin et al. 2018a; 2018b), we take the ten words of both the preceding and following context of the zero pronoun to encode zp, and the context of candidate antecedent is handled with the same way. In addition, we take the last eight words of content of candidate antecedents when they are more than eight words.

Comparison to Baselines

We report the experiment results (F-score) of HAN-PL and the baselines in Table 2, with the overall results on the complete test dataset and also the results for each source.

As shown in Table 2, our model HAN-PL achieves 60.2% in overall F-score, which significantly outperforms the best baseline (Yin et al. 2018b) by 2.9%. Besides, we run experiments on different sources of test corpus, as shown in the first six columns. We can observe that our model HAN-PL improves performance in 4 of 6 sources of the dataset. More specifically, our model outperforms the best baseline (Yin et al. 2018b) on all documents in F-score: by 2.4% (source NW), 0.1% (source MZ), 1.1% (source WB), 5.6% (source BN), 2.2% (source BC) and 1.5% (source TC). One

Models	NW (84)	MZ (162)	WB (284)	BN (390)	BC (510)	TC (283)	Overall
Zhao and Ng (2007)	40.5	28.4	40.1	43.1	44.7	42.8	41.5
Chen and Ng (2015)	46.4	39.0	51.8	53.8	49.4	52.7	50.2
Chen and Ng (2016)	48.8	41.5	56.3	55.3	50.8	53.1	52.2
Yin et al. (2016)	50.0	45.0	55.9	53.3	55.3	54.4	53.6
Yin et al. (2017)	48.8	46.3	59.8	58.4	53.2	**54.8**	54.9
Liu et al. (2016)	59.2	51.3	60.5	53.9	55.5	52.9	55.3
Yin et al. (2018a)	63.1	50.2	**63.1**	56.7	57.5	54.0	57.2
Yin et al. (2018b)	64.3	52.5	62.0	58.5	57.6	53.2	57.3
HAN-PL	**66.7**	**52.6**	**63.1**	**64.1**	**59.8**	54.7	**60.2**

Table 2: Experimental results (%) of comparison to baselines. The first six columns show the results on the different source of documents and the last is the overall result. The strongest F-score in each row is in **bold**. The parenthesized number beside a source's name is the number of zero pronouns in that source.

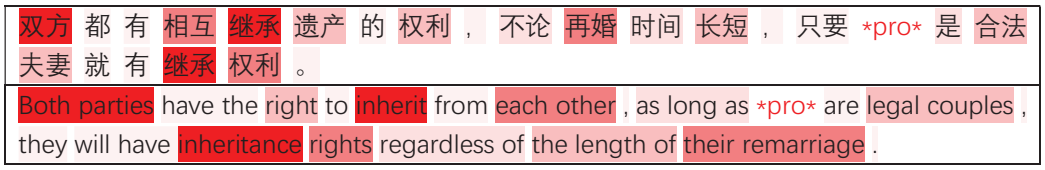

Figure 4: Heat maps of the final self attention weights of HAN. The darker the text background, the greater the attention weight.

Models	Performance
HAN-PL w/o self-attention	58.6
HAN-PL w/o npc2zp attention	59.4
HAN-PL w/o zp2npc attention	58.2
HAN-PL	**60.2**

Table 3: Experimental results (%) on examining the effectiveness of Hierarchical Attention Network.

Models	Performance
HAN-PL w/o pairwise-margin loss	56.5
HAN-PL w/o similarity constraint	58.6
HAN-PL	**60.2**

Table 4: Experimental results (%) on examining the effectiveness of Pairwise Loss.

of the reasons why our model gains better performance on some sources (NW, BN, BC) than the other ones (MZ, WB, TC) may be the short length of text in the latter sources, which makes attention mechanism hard to capture informative information. In addition, some common numerous verbose words such as "Er" and "Yo" in these sources also bring difficulties to encode zero pronouns and candidates.

Effectiveness of Hierarchical Attention

To verify the effectiveness of using hierarchical attention, we conduct extensive experiments on the OntoNotes 5.0 dataset and report experimental results, as shown in Table 3. We design three ablated models:

- **HAN-PL w/o self-attention** applies the bidirectional attention layer, and uses mean pooling on the final outputs from the bi-attention layer, namely H'_{zp} and H'_{npc};
- **HAN-PL w/o npc2zp attention** doesn't apply attention mechanism from npc to zp, and the self attention layer is applied on H_{zp} and H'_{npc};
- **HAN-PL w/o zp2npc attention** is similar to the last model, and the self attention layer is applied on H'_{zp} and H_{npc}

From the experimental results, we can see both the bidirectional attention layer and the self attention layer can get improved performances. Without self attention mechanism, the performance is 1.6% lower than the original method. In addition, changing the bidirectional attention layer to unidirectional attention layers, namely only applying npc2zp attention or zp2npc attention, will also make worse performance. Actually, applying attention from zp to npc is more effective, since the performance of the corresponding ablated model is much worse than that of the original method.

To better illustrate the effectiveness of the Hierarchical Attention Network, we give a case study, as is shown in Figure 4. From the figure, we can see that hierarchical attention learning between zero pronouns and candidate antecedents can successfully capture useful information for explaining the zero pronoun "*pro*" and the candidate antecedent "both parties". The context "inheritance rights" of zero pronoun and the context "the right to inherit" of antecedent, which have similar meanings, get more attention, while some meaningless words, like "from", "are" and some punctuation, are ignored by the Hierarchical Attention Network. Finally the model can achieve the result that the zero pronoun and the candidate antecedent are coreference.

Examples	AttentionZP (Yin et al. 2018b)	HAN-PL
The local people have to cultivate the land. We saw a lot of land being reclaimed (by) *pro* all the way.	The local people have to cultivate the land. **We** (✗) saw a lot of land being reclaimed (by) *pro* all the way.	**The local people** (✓) have to cultivate the land. We saw a lot of land being reclaimed (by) *pro* all the way.
We all know that we need biological diversity, why don't *pro* need cultural diversity?	We all know that we need **biological diversity** (✗), why don't *pro* need cultural diversity?	**We** (✓) all know that we need biological diversity, why don't *pro* need cultural diversity?
China doesn't have any reconnaissance satellites. If *pro* does, it should send back photos at once.	China doesn't have any **reconnaissance satellites** (✗). If *pro* does, it should send back photos at once.	**China** (✓) doesn't have any reconnaissance satellites. If *pro* does, it should send back photos at once.

Table 5: Case study. "*pro*" denotes the position of zero pronouns, and the words marked in bold are the antecedents predicted by the models. Correct and incorrect predictions are marked with ✓and ✗, respectively.

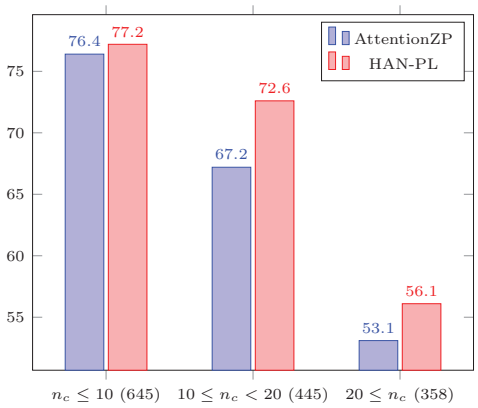

Figure 5: Experimental Results (%) on OntoNotes w.r.t different size of candidate set $n_c = \|S_{zp}\|$. The parenthesized number is the number of zero pronouns whose correct antecedents are in the candidate set.

Effectiveness of Pairwise Loss

To verify the impact of applying Pairwise Loss, we conduct extensive experiments on the OntoNotes 5.0 dataset and experimental results are shown in Table 4. There are two ablated models designed:

- **HAN-PL w/o pairwise-margin loss** changes the final output layer to a softmax layer, and use the cross entropy loss to guide the training of the model, which is applied in previous methods (Yin et al. 2018b; Chen and Ng 2016; Zhao and Ng 2007);
- **HAN-PL w/o similarity constraint** doesn't consider L_c, namely λ_c is set to 0;

As the results show, Pairwise Loss is a reasonable and effective method to guide the optimization of the model. Replacing pairwise-margin loss with cross entropy loss, the performance falls sharply by 3.7%, which confirms that applying pairwise-margin loss is crucial for good performance. Moreover, similarity constraint among np representations of S_{zp}^T, which can utilize the global information between correct candidates, is also helpful for improving the performance according to the experimental results.

Effects of Number of Candidates

To examine how the size of the candidate set effects on the performance of the model HAN-PL, we conduct extensive experiments and the results are shown in Figure 5. As the figure shown, the more candidates there are, the harder it is to find the correct antecedents. However, our model have a more significant improvement compared to AttentionZP (Yin et al. 2018b) when the candidate set becomes larger, for the reason that the attention network we designed can generate more powerful features and the pairwise-loss can make the model more discriminative.

Case Study

Table 5 shows some qualitative cases sampled from HAN-PL and AttentionZP (Yin et al. 2018b). We can observe that our model HAN-PL can perform better for examples which are more complicated. Take the third sample as example, the candidate "China" and "reconnaissance satellites" have similar contexts, so it's hard to distinguish which one is the correct antecedent. However, our model can learn the difference between these two candidates with interaction attention and pairwise-margin loss. Therefore, the cases shows the effectiveness of our model on the task of zero pronoun resolution.

Conclusion

In this paper, we propose an effective model of Hierarchical Attention Network with Pairwise Loss for Chinese zero pronoun resolution. We design a two-layer attention model to better model both zero pronouns and candidate antecedents. To guide the training of the model in a more reasonable way, we also integrate constraint of similarities among correct antecedents into pairwise-margin loss. The experiments on the OntoNotes 5.0 dataset clearly show that the performance of our model is state-of-the-art.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Grant no.61772568), the Natural Science Foundation of Guangdong province (Grant no.2019A1515012029), the Guangzhou Science and Technology Program (Grant no. 201804010288), and the Fundamental Research Funds for the Central Universities (Grant no.18lgzd15).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chen, C., and Ng, V. 2013. Chinese zero pronoun resolution: Some recent advances. In *EMNLP*, 1360–1365.

Chen, C., and Ng, V. 2014. Chinese zero pronoun resolution: An unsupervised approach combining ranking and integer linear programming. In *AAAI*, 1622–1628.

Chen, C., and Ng, V. 2015. Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-the-art resolvers. In *ACL*, volume 2, 320–326.

Chen, C., and Ng, V. 2016. Chinese zero pronoun resolution with deep neural networks. In *ACL*, volume 1, 778–788.

Clark, K., and Manning, C. D. 2016. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Converse, S. P., and Palmer, M. S. 2006. *Pronominal anaphora resolution in Chinese*. Citeseer.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Kadlec, R.; Schmid, M.; Bajgar, O.; and Kleindienst, J. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, F., and Ng, H. T. 2013. Exploiting zero pronouns to improve chinese coreference resolution. In *EMNLP*, 278–288.

Kong, F., and Zhou, G. 2010. A tree kernel-based unified framework for chinese zero anaphora resolution. In *EMNLP*, 882–891.

Liu, T.; Cui, Y.; Yin, Q.; Zhang, W.; Wang, S.; and Hu, G. 2016. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. *arXiv preprint arXiv:1606.01603*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *CVPR*, 1386–1393.

Wang, W.; Yan, M.; and Wu, C. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. *arXiv preprint arXiv:1811.11934*.

Wiseman, S. J.; Rush, A. M.; Shieber, S. M.; and Weston, J. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. Association for Computational Linguistics.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *NAACL*, 1480–1489.

Yeh, C.-L., and Chen, Y.-C. 2007. Zero anaphora resolution in chinese with shallow parsing. *Journal of Chinese Language and Computing* 17(1):41–56.

Yin, Q.; Zhang, W.; Zhang, Y.; and Liu, T. 2016. A deep neural network for chinese zero pronoun resolution. *arXiv preprint arXiv:1604.05800*.

Yin, Q.; Zhang, Y.; Zhang, W.; and Liu, T. 2017. Chinese zero pronoun resolution with deep memory network. In *EMNLP*, 1309–1318.

Yin, Q.; Zhang, Y.; Zhang, W.; Liu, T.; and Wang, W. Y. 2018a. Deep reinforcement learning for chinese zero pronoun resolution. *arXiv preprint arXiv:1806.03711*.

Yin, Q.; Zhang, Y.; Zhang, W.; Liu, T.; and Wang, W. Y. 2018b. Zero pronoun resolution with attention-based neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, 13–23.

Zhao, S., and Ng, H. T. 2007. Identification and resolution of chinese zero pronouns: A machine learning approach. In *EMNLP*.