Zero Pronoun Resolution with Attention-based Neural Network

Qingyu Yin[♯], Yu Zhang[♯]*, Weinan Zhang[♯], Ting Liu[♯], William Yang Wang[♭]
[♯]Harbin Institute of Technology, China
[♭]University of California, Santa Barbara, USA
{qyyin, yzhang, wnzhang, tliu}@ir.hit.edu.cn
william@cs.ucsb.edu

Abstract

Recent neural network methods for zero pronoun resolution explore multiple models for generating representation vectors for zero pronouns and their candidate antecedents. Typically, contextual information is utilized to encode the zero pronouns since they are simply gaps that contain no actual content. To better utilize contexts of the zero pronouns, we here introduce the self-attention mechanism for encoding zero pronouns. With the help of the multiple hops of attention, our model is able to focus on some informative parts of the associated texts and therefore produces an efficient way of encoding the zero pronouns. In addition, an attention-based recurrent neural network is proposed for encoding candidate antecedents by their contents. Experiment results are encouraging: our proposed attention-based model gains the best performance on the Chinese portion of the OntoNotes corpus, substantially surpasses existing Chinese zero pronoun resolution baseline systems.

Title and Abstract in Chinese

基于注意力神经网络模型的零指代消解研究

传统的关于零指代的方法提出了多种关于先行语和零代词的表示模型。这些模型中，研究者们用零代词的上下文信息来帮助表示缺省的信息。为了更好的帮助建模零代词，我们提出了一种基于注意力机制的神经网络模型，通过注意力模型来获取更有表示性信息的上下文信息。实验结果表明：我们的方法能够有效提升效果，并在中文OntoNotes 5.0数据集上取得了最好的结果，超越了现有的基准系统。

1 Introduction

In natural languages, expressions that can be deduced contextually by people are frequently omitted in texts. This is special the case in pro-dropped languages, such as Chinese, where a kind of anaphoric expression is frequently eliminated. A zero pronoun is a gap in the sentence that is found when a phonetically null form is used to refer to a real-world entity (Chen and Ng, 2016). We here show a case of zero pronouns from the OntoNotes-5.0 dataset.

> 这 次 地震 ϕ_1 有 一些 房屋 塌 的 , 这 里面 如果 有 建房 的 质量 问题 ， ϕ_2 是 要 追究 责任 的 。
>
> In this earthquck ϕ_1 some rooms collapsed, if there exsit some room quality issues, ϕ_2 will need to call to account.

We use ϕ to represent the zero pronouns in this example. Among these zero anaphoras, we can assign the mention "政府/the government" that appears in leading text, to be the antecedent of ϕ_2 while there are no such mentions for ϕ_1. Hence, ϕ_2 is an anaphoric zero pronoun, and ϕ_1 is the un-anaphoric case.

*Corresponding author.

Proceedings of the 27th International Conference on Computational Linguistics, pages 13–23
Santa Fe, New Mexico, USA, August 20-26, 2018.

With the fact that zero pronouns are gaps that have no text, it is almost impracticable to representant the zero pronouns by themselves. This issue has received increasingly attention. In recent time, deep neural network methods for Chinese zero pronoun resolution (Chen and Ng, 2016; Yin et al., 2017a; Yin et al., 2017b) have been proposed and are intended to encode zero pronouns into the vector-space semantic by additional elements. Chen and Ng (2016) propose a neural network model that learns to encode the anaphoric zero pronoun by using the leading word and governing verb, which leads to the insufficient text issue. To better use associated text information for expressing zero pronouns, Yin et al. (2017a) present the ZP-centered-LSTM architecture that learns to encode zero pronouns by their text words. However, it could bring with some defects: their model regards all the words in the sentence equally, thus fails in capturing informative parts of the sentence. As described in (Chen and Ng, 2016), the clause information of a zero anaphora is very important in explaining these gaps, it is a natural way of modeling zero pronoun by focusing on some important texts. For instance, in sentence "two trains in ten weeks ago ϕ crashed in their way to the southern German countryside", "crashed in their way" is an important clue for the zero pronoun but "several weeks ago" is not. Under this consideration, a systematic solution that can encode zero pronouns by focusing on informative parts of associate texts is the better choice. Besides, another important issue for the task of zero pronoun resolution is modeling candidates. Recent researches use context and content words to encode candidates (Yin et al., 2017a; Yin et al., 2017b). Typically, these words are modeled in a sequential way by recurrent neural networks. We argue that some of the words in noun phrases contains more important information than others, a need that leads to the usage of attention mechanism.

To alleviate the above-mentioned issues, in this paper, we propose a novel attention-based neural network model to deal with the task. Following existing neural network work for Chinese zero pronoun resolution (Chen and Ng, 2016; Yin et al., 2017a; Yin et al., 2017b), we focus on anaphoric zero pronoun resolution task, introducing a pair-wise model to resolve anaphoric zero pronouns. For some natural language processing tasks (Mnih et al., 2014; Tang et al., 2016), people investigate to apply attention mechanism on top of the convolutional neural network or recurrent neural network to introduce an extra source of information to guide the modeling of useful information. However, since zero pronouns are simple gaps that have no such kind of extra information, the above-mentioned attention mechanism can rarely be directly practiced for modeling these gaps. Inspired by (Lin et al., 2017), we here investigate the usage of a self-attentive mechanism for encoding the zero anaphoras. With the help of self-attentive mechanism, our model is able to effectively focus on informative texts of the zero pronouns and therefore captures essential information on encoding the zero pronouns. In addition, on purpose of modeling informative texts of mentions (noun phrases), we propose an attention-based recurrent neural network to build the mention encoder. With the help of representative vector of zero anaphoras, our model is able to effectively focus on important parts of mentions and therefore brings an efficient way of expressing candidates at the semantic level. Empirically, we show that our method has brought performance gains in baselines, achieving great performance on the widely used OntoNotes-5.0 dataset. Our contributions are three-fold:

- By utilizing the self-attention mechanism, our model is able to focus on informative parts of associate texts when modeling zero pronouns, leading to an effective way of capturing useful information for interpreting the zero pronouns;

- Our model is capable of modeling candidate antecedents by their informative words with the help of zero pronouns, which in return brings a better way of explaining candidate antecedents;

- We show that our model substantially surpasses all baseline systems, gains state-of-the-art performance on the benchmark dataset.

In Section 2, we will discuss related work on zero pronoun resolution. Next, we will intorduce our attention-based neural network model in Section 3. In Section 4, empirical evaluation results are shown. And finally, we conclude in Section 5.

2 Related Work

In this section, we give a brief summary of early efforts for zero pronoun resolution both for Chinese and other languages.

2.1 Zero Pronoun Resolution for Chinese

Converse (2006) is the first rule-based study that integrate Hobbs-algorithm into the resolution of zero pronoun in the Chinese Treebank (Xue et al., 2005). After that, a variety of learning-based methods have been investigated. Zhao and Ng (2007) use the learning-based model to locate and resolve zero anaphoras. They investigate a serious of features and apply the decision-tree algorithm to train the classifier. To better capture the syntactic-level information, Kong and Zhou (2010) introduce the context sensitive tree-kernel unified framework for zero anaphor resolution. On the base of Zhao and Ng (2007), Chen and Ng (2013) further investigate their model, introducing two extensions to the resolver, namely, novel features and zero pronoun links. However, these work deeply rely on annotation dataset. To alleviate this issue, Chen and Ng (2014) present the first unsupervised model that first convert zero anaphoras into ten pre-defined pronouns and then apply a ranking-based pronoun resolution model to select antecedent mentions. Chen and Ng (2015) build a discourse-aware model that can jointly locate and resolve zero anaphoras.

More recently, with the advance of neural network techniques, deep-learning-based methods are introduced and have been demonstrated to be effective for this task. Chen and Ng (2016) first introduce a feed-forward neural network framework, where zero anaphoras are encoded by its previous word and headword. However, their model overlooks context of a zero anaphora, which inevitably misses some valuable information. Naturally, some works try to alleviate this issue by investigating information from associate texts. Yin et al. (2017a) introduce a novel memory-based network neural network model that learns to encode zero anaphoras by its texts and antecedent mentions. They take advantage of multi-hops architecture, producing abstract information from external-memories as hints for explaining zero anaphoras. Yin et al. (2017b) focus on encoding global-information for candidates, where a hierarchical candidate encoder is introduced that learns to model the candidates. Liu et al. (2017) investigate the issue of generating pseudo training-data for the task of zero anaphora resolution. They use a novelty two-step training strategy that helps to overcome the diversity between the generated pseudo training-data and the real one. Even though these above-mentioned methods can reveal the semantic of zero anaphoras by its context, they regard all the words equally, overlooking the diversity of different words. In this paper, we focus on exploring an effective way of modeling zero pronoun by using the associated texts. More specifically, we integrate a novel self-attentive mechanism, which provides our model an ability to focus on multi-aspects text, benefiting the encodings of zero anaphoras. In addition, by employing an attention-based technique for modeling candidates, our model learns to encode more informative parts of the mentions. All these bring advantages to the resolution of zero pronouns.

2.2 Zero Pronoun Resolution for other Languages

There has been a variety of work on zero pronoun resolution for other languages besides Chinese, such as Korean and Japanese. These methods could be categorized as rule-based and learning-based. Ferrández and Peral (2000) investigate a rule-based method that can encode preferences for candidates for resolving zero anaphoras in Spanish. In recent time, learning-based methods (Han, 2006; Iida and Poesio, 2011; Isozaki and Hirao, 2003; Iida et al., 2006; Iida et al., 2007; Sasano and Kurohashi, 2011; Iida and Poesio, 2011; Iida et al., 2015; Iida et al., 2016) have been well studied. Iida et al. (2016) present a novel CNN-based deep neural network model for intrasentential subjective zero anaphora resolution in Japanese. As clues, they use both the surface-word and the dependency tree-structure of a sentence. Their model gains higher precision, which is needed for real-world natural language processing (NLP) applications.

3 Methodology

We introduce an attention-based neural network model for anaphoric zero pronoun resolution. Compared to the prior studies that have the underutilized context of zero pronouns, we investigate an attention

mechanism that helps to effectively capture useful information from associate texts. We here present the methodology in details, which include the preliminary, the architecture of proposed attention-based neural network and training objective of the model.

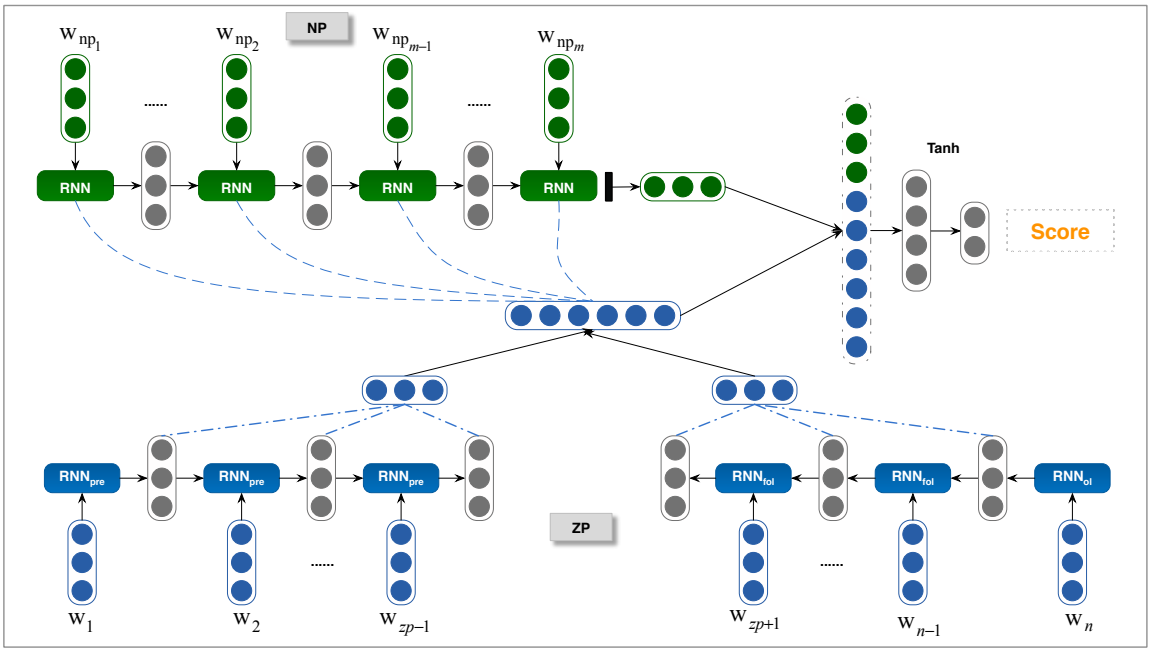

Figure 1: Framework of the proposed attention-based neural network for zero pronoun resolution. The w_i for zero pronoun part means the i-th word in the associated sentence, w_{zp-i} is the i-th word before the zero pronoun and w_{zp+i} is the i-th word behind the zero pronoun. w_{np_i} is the i-th word of the noun phrase. After generating the representative vector of zero pronoun and candidate mention, we generate its resolution score by going through two $tanh$ layers.

3.1 Preliminary

In the first place, we select its candidate mentions from the associated texts. More specifically, we choose the noun phrases that appear within two sentences from the zero anaphora to be the candidates. In addition, we choose the strategy used in prior approaches (Chen and Ng, 2016; Yin et al., 2017a) for Chinese zero pronoun resolution, reserving those mentions that are max noun phrases or modifier noun phrases as candidates. By doing so, we can recall most (about 98%) of the antecedents with a small loss. Then, what we desire the resolver to do is to accurately recognize antecedents for zp from its candidates list $NP = \{np_1, np_2, ..., np_n\}$.

Our model is basically a pair-wise model (Zhao and Ng, 2007; Chen and Ng, 2016; Yin et al., 2017b). For each candidate mention of zp, we classify it into two classifications, namely, "corefer" that means the candidate is the antecedent of the zero pronoun; or otherwise, "un-corefer". We build the classifier by applying the attention-based techniques. More specifically, an attention-based neural network model is utilized that generates the coreference score of each zero pronoun-candidate antecedent pair.

3.2 Attention-based Neural Network Model

In this part, we introduce our attention-based neural network for anaphoric zero pronoun resolution in detail. The architecture of our model is shown in Figure 1. For an anaphoric zero pronoun zp, we convert the input instances into real-valued vectors by using its context words. With the help of the self-attentive mechanism, our model learns to represent zero pronouns by focusing on different parts of contexts. In this way, we generate the representative vector of zp as v_{zp}. In addition, when dealing with candidate mentions, our model learns to encode the candidate mentions by using its informative content words. We here manipulate the v_{zp} as an external vector to attend to the informative parts of its candidate mentions.

By doing so, our model captures the important information of each candidate according to the zp. After that, we generate the representative vectors of candidates as $\{v_{np_1}, v_{np_2}, ..., v_{np_n}\}$. Lastly, we feed the representative vectors of the candidate and zero pronoun to a two-layer neural network, generating the resolution score for each zero pronoun-candidate antecedent pair. After that, we obtain the resolution probability for each candidate. The candidate mention with the biggest probability is regarded as the final result. Basically, our model involves three modules, namely, a zero pronoun encoder; a candidate antecedent encoder; and a feed-forward neural network that learns to score each candidate antecedent.

3.2.1 Modeling Zero Pronoun

Inspired by Lin et al. (2017), we investigate the self-attention mechanism when modeling zero anaphoras. In practice, we use two recurrent neural networks (RNNs) to encode the preceding and following texts of the zero anaphora (Yin et al., 2017a). The last hidden vectors of these two RNNs are concatenated as the vector-space semantic of the zero pronoun. On top of the employed recurrent neural network architecture, we apply the proposed attention technique that helps the model to capture the informative parts of associated texts.

Especially, our self-attention mechanism provides the attention-weight vectors for hidden states of the employed RNN architectures. We then dot these attention-weight vectors with counterpart hidden states and use the weighted summation vector as the representative vector for the zero pronoun. For a zero pronoun, we map its associated text words as sequential embeddings.

$$Context_{precedding} = (w_1, w_2, ..., w_{zp-1}) \tag{1}$$

$$Context_{following} = (w_{zp+1}, w_{zp+2}, ..., w_n) \tag{2}$$

where w_i is a d dimensional embedding for the ith word in the sentence. After that, we generate the representative vector of the preceding and the following text by using two separate RNNs:

$$h_t^{pre} = RNN_{pre}(w_t, h_{t-1}^{pre}) \tag{3}$$

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

where RNN_{pre} and RNN_{fol} are two employed RNNs that model the precedding and following context of the zero pronoun independently. After that, we get the hiddent vector for each word, which has the dimension of u. We represent all the hidden states of RNN_{pre} and RNN_{fol} as $H_{pre} \in \mathbb{R}^{n_{pre} \times u}$ and $H_{fol} \in \mathbb{R}^{n_{flo} \times u}$, seperately:

$$H_{pre} = \{h_1^{pre}, h_2^{pre}, ..., h_{n_{pre}}^{pre}\} \tag{5}$$

$$H_{fol} = \{h_1^{fol}, h_2^{fol}, ..., h_{n_{fol}}^{fol}\} \tag{6}$$

We then apply the self-attention mechanism, which computes linear-combinations of the hidden vectors in H_{pre} and H_{fol}. The attention mechanism takes H_{pre} (or H_{fol}) as the inputs and produces a matrix of attention-weight A_{pre} (A_{fol}):

$$A_{pre} = softmax(W_2^{pre} tanh(W_1^{pre} H_{pre}^T)) \tag{7}$$

$$A_{fol} = softmax(W_2^{fol} tanh(W_1^{fol} H_{fol}^T)) \tag{8}$$

where W_1 is a weight matrix with a shape of d_a-by-u and W_2 is in shape of r-by-d_a; r represents the number of hops of attention we choose. The $softmax()$ is performed along the second dimension of its input. In this way, the attention matrix A could be seen as a multi-hope attention matrix. Comparing with the single-attention matrix, such a mechanism enables our model to focus on different parts of the contexts, bringing a more efficient way of modeling sentence-level information for the zero pronoun at the semantic level.

We get the r weighted sums by multiplying the attention matrix A and hidden states H, regarding the resulting matrix as the representative vector of the zero pronoun's preceding and following texts:

$$M_{pre} = A_{pre} H_{pre} \tag{9}$$

$$M_{fol} = A_{fol} H_{fol} \tag{10}$$

Subsequently, we obtain the representative vectors of the associated text of the zero pronoun by averaging the row vectors in each representative matrix (namely, M_{pre} and M_{fol}). After that, these two vectors are concatenated as the representative vector of the zero pronoun. Our experiments show that the attention-based model leads to a considerable improvement from the non-attentive model, indicating that the self-attention mechanism can help to better encode the zero pronoun, focusing on informative parts of the associated texts.

3.2.2 Modeling Candidate Antecedent

We here build the candidate antecedent encoder by using an RNN architecture, whose input is comprised by the words in the candidate antecedent. In an effort to better align the more informative parts of phrases to the anaphoric zero pronoun, we here integrate an attention technique into our model. In this work, we use a gating-function as our attention mechanism. Especially, given the representative vector of the anaphoric zero pronoun $v^{(zp)}$, the output vector and input embedding vector of RNN in the candidate mention part at time t, $h_t^{(np)}$ and e_t, the attention mechanism computes a gate as: $attention_t = att(e_t, h_t^{(np)}, v^{(zp)})$, where att is defined as:

$$s_t = tanh(W^{(att)} \cdot [e_t; h_t^{(np)}; v^{(zp)}] + b^{(att)}) \tag{11}$$

$$attention_t = \frac{exp(s_t)}{\sum_{t'=1}^{m} exp(s_{t'})} \tag{12}$$

where $W^{(att)}$ and $b^{(att)}$ are parameters to be learned, m is the number of words in the mention. After that, we regard the averaged attention-hidden vector as the vector-space semantic of the candidate antecedent, which takes considerations of a hierarchy of historical semantic:

$$\tilde{v}_{np} = \sum_{i=1}^{m} h_i^{(np)} * attention_i \tag{13}$$

3.2.3 Calculating Resolution Scores

After generating the representive vector of zero pronoun, v_{zp} and vectors of its candidates $\{\tilde{v}_{np1}, \tilde{v}_{np2}, ..., \tilde{v}_{npn}\}$, we calcualte the resolution score for each zero pronoun-candidate antecedent by using a two-layers feed-forwd nerual network. Taking v_{zp} and its i-th candidate mention v_{np_i} as inputs, our model calcuate the resolution score by going through two $tanh$ layers:

$$s_j = tanh(W_i^{(s)} \cdot s_{j-1} + b_j^{(s)}) \tag{14}$$

where $s_0 = [v(zp); v(np_i); v_i^{(fe)}]$, $W^{(s)}$ and $b^{(s)}$ are the parameters of this feed-forward neural network. In addition, to better capture the syntactics, position and other relations between an anaphoric zero pronoun and its candidates, we encode hand crafted features ($v^{(fe)}$) as inputs to our neural network model. We utilize the features from exsiting work on zero anaphora resolution (Chen and Ng, 2013; Chen and Ng, 2016), map them into vectors to estimate the resolution score for the zero pronoun-candidate mention pair as:

$$score_i = W^{(sco)} \cdot s_{-1} + b^{(sco)} \tag{15}$$

where $score_i$ denotes the probability of the i-th candidate mention (np_i) being predicted to be the antecedent, and s_{-1} is the output vector of the second hidden layer. After that, we obtain the resolution scores for all the candidates $\{score_1, score_2, ..., score_n\}$. The candidate mention with the biggest score is eventually selected to be the antecedent of the anaphoric zero pronoun.

3.3 Training Objective

Same as Yin et al. (2017b), we train our model by minimizing the cross entropy error of coreference classification. The training objective is defined as:

$$loss = -\sum_{t \in T} \sum_{np \in NP} \delta(zp, np) \log(P(zp, np)) \tag{16}$$

where T represents all training instances, NP is the candidate-set of the anaphoric zero pronoun zp; $\delta(zp, np)$ represents the coreference of zp and its candidate mention np: if they are coreference, $\delta(zp, np) = 1$ or otherwise, $\delta(zp, np) = 0$.

4 Experiments

4.1 Experiment Setup

4.1.1 Evaluation Metrics

Same as early work on Chinese zero pronoun resolution (Zhao and Ng, 2007; Chen and Ng, 2016; Yin et al., 2017a; Yin et al., 2017b), we manipulate to evaluate the quality of our model by Recall, Precision and F-score (denoted as F). More specifically, recall and precision are defined as:

$$Recall_{Res} = \frac{\#\ Res\ Hit}{\#\ AZP\ in\ Key} \tag{17}$$

$$Precision_{Res} = \frac{\#\ Res\ Hit}{\#\ AZP\ in\ Predictions} \tag{18}$$

where a "Res Hit" means that the anaphoric zero pronoun is successfully identified and successfully resolved to a candidate mention that is in the same coreference chain as in the golden answer key annotated in the dataset.

4.1.2 Experiment Settings

Same to existing work on Chinese zero pronoun resolution (Chen and Ng, 2016; Yin et al., 2017a; Yin et al., 2017b), we run experiments on the Chinese part of the OntoNotes-5.0 dataset[1] used in the Conll-2012 task. Because zero pronoun coreferences are only annotated in the training and development set, we thus train our model on the training dataset and evaluate the model on the development dataset. Table 1 is the statistics of our dataset.

	Documents	Sentences	Words	Anaphoric Zero Pronouns
Training	1,391	36,487	756K	12,111
Test	172	6,083	110K	1,713

Table 1: Statistics on the training and test dataset.

We use the recent zero pronoun resolution systems 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; Liu et al., 2017; Yin et al., 2017a; Yin et al., 2017b). As we are focusing on the anaphoric zero pronoun resolution, we run experiments by directly employing golden parse tree and golden anaphoric zero pronouns that are annotated in the dataset. In addition, documents in the datasets are from 6 sources: **BN** (Broadcast News), **NW** (Newswire), **BC** (Broadcast Conversation), **WB** (Web Blog), **TC** (Telephone Conversation) and **MZ** (Magazine). We report the overall results on the complete test dataset and also the result from the different source of the dataset.

[1] `http://catalog.ldc.upenn.edu/LDC2013T19`

	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	55.4	50.8	53.1	52.2
Yin et al. (2017b)	50.0	45.0	55.9	53.3	55.3	54.4	53.6
Yin et al. (2017a)	48.8	46.3	59.8	58.4	53.2	**54.8**	54.9
Liu et al. (2017)	59.2	51.3	60.5	53.9	55.5	52.9	55.3
Our model	**64.3**	**52.5**	**62.0**	**58.5**	**57.6**	53.2	**57.3**

Table 2: Experiment results on the test dataset, including the results on the overall dataset and different sources of the dataset. 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 anaphoric zero pronouns in that source.

4.2 Hyperparameter

To tune the hyperparameters of our model, 20% of the training dataset are reserved as a held out development set. Such a strategy is also utilized in the baseline systems (Chen and Ng, 2016; Yin et al., 2017a). We randomly initialize the parameters and minimize the loss-function by Adagrad (Duchi et al., 2011) with learning-rate 0.003. The input embedding vector dimension is 100, the dimension of hidden layer of RNNs (namely, the u) is 256 and d_a is 128. Besides, we fix the dimensions of two hidden layer of the feed-forward neural network to 256 and 512. We add the dropout (Hinton et al., 2012) with a probability of 50% on the output of each layer. The code for this work is released in `https://github.com/qyyin/AttentionZP.git`.

4.3 Experiment Results

We report the experiment results (F-score) of our model and the baselines in Table 2. The number of hops of attention is fixed to be 2 ($r = 2$), where we gain the best result. We report the overall results on the complete test dataset and also the results for each source of documents. As we can observe that our model gains 57.3% in overall F-score, which significantly beats the best baseline system (Liu et al., 2017) by 2.0%. In addition, we run experiments on different sources of test corpus, as shown in the first six columns Table 2. The parenthesized number beside a source's name represents the number of anaphoric zero pronouns in that source. We can observe that our model improves performance significantly in 5 of 6 sources of the dataset. More specifically, our model beats the best baseline (Liu et al., 2017) on all documents in F-score: by 5.1% (source NW), 1.2% (source MZ), 1.5% (source WB), 4.6% (source BN), 2.1% (source BC) and 0.3% (source TC). The main reason why our model gains worse performance on source "TC" lie in the short length of text in this source, which makes our model hard to learn to capture useful information for expressing the zero pronouns. Besides, there are full of numerous verbose words such as "呃/Er", "哟/Yo", which brings difficulties for our model to accurately encode a zero pronoun by focusing on its informative contexts. More efforts could be performed in order to encode the associated text of a zero pronoun in a more efficient way, modeling word-sequences beyond the sentence boundary, for instance.

We show in Figure 2 the learning curve of our model on the development dataset. As we can observe, after the first epoch, the F-score on the test dataset is about 46%, and it gradually grows to 52% after about 30 iterations when performance starts to plateau. It is well accepted that modeling useful parts of associated text play an important role in encoding the zero pronouns. By applying the self-attentive mechanism, our model learns to focus on important parts of the contexts, revealing multi-aspect sentence-level information in a more efficient way than those without attention. On the other side, by assigning different attention to the words in a candidate mention with respect to the information of the counterpart zero pronoun, our model learns to encode candidate mentions in a more natural way. Hence, both of them bring benefit to selecting accurate antecedents, leading to better performance.

In addition, having multi-aspect sentence-level information is expected to afford more abundant in-

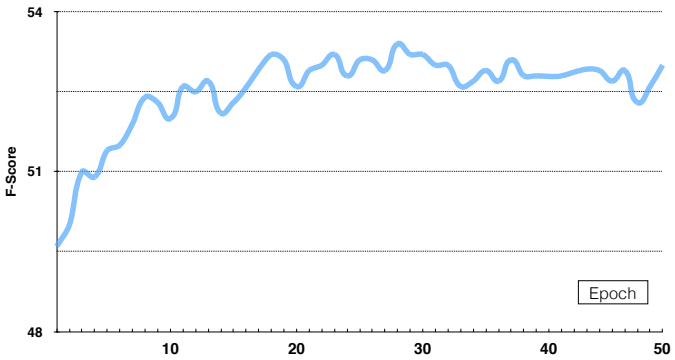

Figure 2: Learning curve of our model on the development dataset.

formation about the encoded zero pronoun, we thus evaluate how the improvement can be brought by tuning r in the self-attentive mechanism. We vary r from 2 to 6, as is shown in Figure3. We can observe that the best performance is reached when $r = 2$. The results are not confusing because we are tried to focus on the informative part for zero pronouns, and we cut the sentence into two separate parts that are zero pronoun-centric, when $r = 2$ means to attend on totally 4 parts of the sentence, thus our model performances well in this situation.

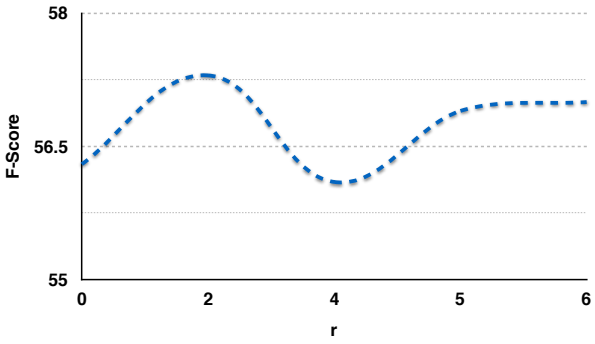

Figure 3: Effect of tuning r for encoding zero pronouns.

For illustrating the effect of the proposed attention mechanism for zero pronoun and candidate mentions, we also run an experiment without applying the attention mechanism for both zero pronouns and candidates, which is $r = 0$ in Figure 3. As we can observe that the performance drops by removing the attention mechanism. For the model without the attentive mechanism, the performance drops by 0.9% in F-score compared with that of the full model. Because that the proposed self-attentive mechanism for zero pronoun brings our model an ability to access multi-aspect sentence-level information, removing such an architecture unsurprising influences the results significantly. With an inspiration that not all the words are equally important for explaining the mention, the performance of removing the attention for candidate mentions is reasonable weak. All these show the benefit of our attention-based model.

Lastly, we give a case study to illustrate the power of our self-attentive mechanism, as is shown in Figure 4. From the figure, we can tell that the model successfully focuses on informative parts of the texts. Though there are redundancies between different hops of attention, our model can capture useful information for explaining the zero pronoun (denoted as "*pro*" in the picture). Our model learns to focus on the informative words such as "数字证书/digital certificate", which is essential for expressing the zero pronoun.

Figure 4: Heat maps of our attention-based model. In this case, we show the detailed attention weight taken by the attention matrix taken by the attention matrix ($r = 2$). Darker color means higher weight.

5 Conclusion

We proposed a novel attention-based neural network model for Chinese zero pronoun resolution. Using recent advances in attention mechanism, we developed a self-attentive architecture for modeling zero pronouns, which enables our model to focus on parts of the associated texts. In addition, we also investigated an attention-based candidate antecedent encoder that learns to model important parts of the noun phrases with respect to the representative vector of zero anaphoras. Our experiments demonstrated that our method significantly surpasses the state-of-the-art on a benchmark dataset for anaphoric zero pronoun resolution. Future work will evaluate our model on other natural language processing problems, such as anaphora resolution for Chinese and English. We also plan to investigate training the anaphora-specific embedding that could better reveal the descriptive attribute for the zero anaphoras.

Acknowledgments

This work has been supported by the Major State Basic Research Development 973 Program of China (No.2014CB340503), National Natural Science Foundation of China (No.61472105 and No.61502120). We are grateful to Xuxiang, Xiaocheng Feng and anonymous reviewers for their useful feedback. By the meaning by Harbin Institute of Technology, the contact author of this work is Yu Zhang.

References

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

Chen Chen and Vincent Ng. 2014. Chinese zero pronoun resolution: An unsupervised approach combining ranking and integer linear programming. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Chen Chen and Vincent Ng. 2015. Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-the-art resolvers. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, page 320.

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. In *Proceedings of the 54rd Annual Meeting of the ACL*.

Susan P Converse. 2006. Pronominal anaphora resolution in chinese.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Antonio Ferrández and Jesús Peral. 2000. A computational approach to zero-pronouns in spanish. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 166–172. Association for Computational Linguistics.

Na-Rae Han. 2006. *Korean zero pronouns: analysis and resolution*. Ph.D. thesis, Citeseer.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ilp solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 804–813. Association for Computational Linguistics.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 625–632. Association for Computational Linguistics.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4):1.

Ryu Iida, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, and Julien Kloetzer. 2015. Intra-sentential zero anaphora resolution using subject sharing recognition. *Proceedings of EMNLP'15*, pages 2179–2189.

Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. 2016. Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In *Proceedings of EMNLP*.

Hideki Isozaki and Tsutomu Hirao. 2003. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 184–191. Association for Computational Linguistics.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 882–891. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Ting Liu, Yiming Cui, Qingyu Yin, Shijin Wang, Weinan Zhang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In *ACL*.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *IJCNLP*, pages 758–766.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017a. Chinese zero pronoun resolution with deep memory network. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1309–1318.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017b. A deep neural network for chinese zero pronoun resolution. In *IJCAI*.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of chinese zero pronouns: A machine learning approach. In *EMNLP-CoNLL*, volume 2007, pages 541–550.