Tackling Zero Pronoun Resolution and Non-Zero Coreference Resolution Jointly

Shisong Chen[1], Binbin Gu[3], Jianfeng Qu[1], Zhixu Li[2]*,
An Liu[1], Lei Zhao[1], Zhigang Chen[4]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[3]University of California, Irvine [4]IFLYTEK Research, Suzhou, China
sschen777@stu.suda.edu.cn, binbing@uci.edu, zhixuli@fudan.edu.cn
{jfqu,anliu,zhaol}@suda.edu.cn, zgchen@iflytek.com

Abstract

Zero pronoun resolution aims at recognizing dropped pronouns and pointing out their anaphoric mentions, while non-zero coreference resolution targets at clustering mentions referring to the same entity. Existing efforts often deal with the two problems separately regardless of their close essential correlations. In this paper, we investigate the possibility of jointly solving zero pronoun resolution and coreference resolution via a novel end-to-end neural model. Specifically, we design a gap-masked self-attention model that encodes gaps and tokens in the same space, where gaps could capture valuable contextual information according to their surrounding tokens while tokens could maintain original sequential information without disturbance. Additionally, we also propose a two-stage interaction mechanism to make full use of the exclusive relationship between zero pronouns and mentions. Our empirical study conducted on the OntoNotes 5.0 Chinese dataset shows that our model could outperform corresponding state-of-the-art approaches on both tasks.

1 Introduction

Zero pronoun resolution and non-zero coreference resolution are two fundamental tasks in natural language processing (NLP). Zero pronoun resolution, which is only studied in pro-drop languages such as Chinese, aims at recognizing dropped pronouns in a given text and pointing out their anaphoric mentions within the text (Chen and Ng, 2013). Coreference resolution, which is studied in all languages, targets at clustering mentions referring to the same real-world entity in the text (Sukthanker et al., 2020). Both of the two tasks are vital for many downstream NLP applications including machine translation (Mitkov et al., 1995), information extraction (Zelenko et al., 2004) and text summarization (Steinberger et al., 2007).

Given their importance, both tasks have been studied extensively. For zero pronoun resolution, most previous works assume that positions of zero pronouns are given. They encode zero pronouns and their candidate antecedents by LSTM (Yin et al., 2017), Attention (Yin et al., 2018b; Liu et al., 2017) or BERT (Song et al., 2020; Aloraini and Poesio, 2020), then measure the similarity between embeddings to find out the best antecedents. However, positions of zero pronouns are usually unknown in practice. To solve this problem, some recent works (Yang et al., 2019; Song et al., 2020) treat spaces between tokens as candidate zero pronouns, recognize and resolve zero pronouns based on gap embeddings jointly. For non-zero coreference resolution, the state-of-the-art models belong to an end-to-end paradigm or its variants. In these models, all spans in the text are candidate mentions. After representing spans by highway LSTM (Lee et al., 2018) or transformers (Joshi et al., 2019, 2020), they calculate mention and antecedent scores using feed forward neural network or machine reading comprehension method (Wu et al., 2019), and select the top K mentions and their antecedents with the top antecedent score.

Most existing efforts deal with the two problems separately, and someone also state that it is challenging to combine zero pronoun resolution with the resolution of overt mentions (Aloraini and Poesio, 2020). But according to our observations, there are at least four intuitions motivating us to tackle the two tasks jointly:

- Firstly, with the joint learning of zero pronoun resolution and non-zero coreference tasks, a more robust and universal representations of tokens could be learned and shared for both tasks, which could potentially benefit both tasks.
- Secondly, the coreferred mentions in coreference resolution could enrich each others' contextual information, which may provide some hidden clues for zero pronoun resolution. In Fig. 1,

*Zhixu Li is the corresponding author

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 518–527
November 10–11, 2021.

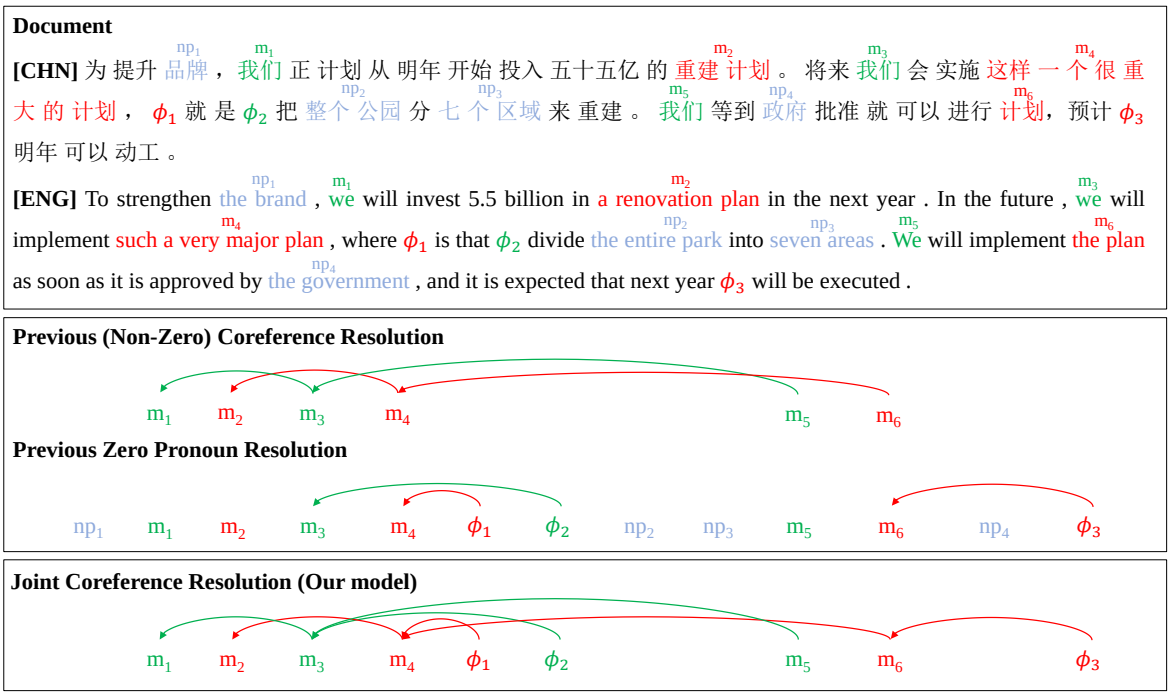

Figure 1: An example document for coreference resolution and zero pronoun resolution, where mentions and zero pronouns marked in the same color have coreferential relationship, and noun phrases that are taken as candidate antecedents of zero pronouns are marked in blue color. For the sake of clarity, we do not list all the mentions, zero pronouns, and noun phrases in the figure.

while resolving, the contextual information of m_2, such as "next year", will be incorporated into the representation of m_6 as they are referring to the same entity. Therefore, ϕ_3 next to "next year" will be explicitly resolved to m_6 with reference to the common context.

- Thirdly, based on the mention detection results in non-zero coreference resolution, only mentions, instead of all maximal and modifier noun phrases in the sentence, would be taken as candidate antecedents. As in Fig. 1, the noun phrase np_1 would not be taken as candidate antecedent in jointly learning.

- Lastly, the positions of zero pronouns and the spans of mentions are exclusive. That is, if a mention is determined, gaps in the mention won't be taken as a zero pronoun. Also, if a zero pronoun is determined, a phrase across the zero pronoun is impossible to become a mention. Therefore, the mention detection results can help to exclude some gaps as zero pronoun, and vice versa. In Fig. 1, if m_4 is a mention, the gaps in the mention, such as the gap between "major" and "plan", should not be a zero pronoun. If ϕ_3 is a zero pronoun, the span "next year will be executed" across the ϕ_3 should not be a mention.

Guided by the four intuitions above, this paper proposes a crafted end-to-end neural model for tackling the two tasks jointly, where the positions of zero pronouns are unknown in advance. Within the model, both tasks share the same token embedding layer (intuition 1). Inspired by Lee et al. (2018) and Zhang et al. (2018), we introduce high-order inference module to enrich mentions' contextual information (intuition 2) and introduce mention detection loss to help mention recognition (intuition 3). Additionally, considering the exclusive relations between the position of gaps and spans, we propose a two-stage interaction mechanism to realize the exclusivity in both embedding level and detection scoring level (intuition 4). Last but not the least, to learn proper representation for both gaps and tokens in the same space, a gap-masked self-attention model is designed which enables gaps to take advantage of the contextual messages from their surrounding tokens, without polluting the representations of tokens.

To summarize, our contributions are as follows:

- We are the first to attempt to solve zero pronoun resolution and coreference resolution with an

end-to-end neural network model jointly.

- We propose a two-stage interaction mechanism to make full use of the exclusive relationship between zero pronouns and mentions.
- We design a gap-masked self-attention model that could learn representations for gaps from their surrounding tokens, without polluting the representations of tokens in the same space.

Extensive experiments conducted on the widely-used OntoNotes 5.0 Chinese dataset shows that our model could outperform the state-of-the-art approaches on both tasks. The code is available at https://github.com/cheniison/e2e-joint-coref.

2 Related Work

In this section, we first briefly introduce previous works on zero pronoun resolution and non-zero coreference resolution respectively, and then introduce some joint works.

According to whether positions of zero pronouns are given, previous works of zero pronoun resolution can be classified into two categories: resolution with gold positions of zero pronouns and resolution without gold positions of zero pronouns.

For resolution with gold positions of zero pronouns, previous work concentrates on finding an antecedent for a given zero pronoun. (Chen and Ng, 2016; Yin et al., 2017) first use deep learning models to compute embeddings for zero pronouns and candidate antecedents. Liu et al. (2017) proposes a method of generating an amount of pseudo zero pronoun data automatically so that models can be pre-trained on the pseudo data. Yin et al. (2018a) introduces reinforcement learning model into zero pronoun resolution to integrate local and global resolution information. Lin and Yang (2020) considers bidirectional attention between zero pronouns and candidate antecedents and proposes a pairwise-margin loss and a similarity constraint to optimize their model.

For resolution without gold positions of zero pronouns, positions of zero pronouns need to be identified first. Kong and Zhou (2010) and Bouzid and Zribi (2020) first find the positions of zero pronouns, after that they filter out non-referential pronouns and determine antecedents. In the work of Chen and Ng (2013), zero pronouns can establish coreference relation with each other to help zero pronouns find overt antecedents far away from them. These works rely on handcraft features and suffer from error propagation. To solve these problems, Yang et al. (2019) tackles context reconstruction in an end-to-end way. It formulates pronouns detection as a sequence labeling task and uses pronoun masking mechanism to combine the detection and the resolution modules. Song et al. (2020) presents a Bert-based multi-task model which handles zero pronoun recovery, zero pronoun detection and zero pronoun resolution jointly.

Non-zero coreference resolution is an important task in natural language processing in all languages. Recently, end-to-end coreference resolution models achieve the state-of-the-art performance in the non-zero coreference task. Lee et al. (2017) proposes the first neural end-to-end coreference resolution model which computes scores of span and mention pairs jointly to detect mentions and predict antecedents. To avoid global inconsistency, Lee et al. (2018) proposes a higher-order model which iteratively updates mention representations. Zhang et al. (2018) learns mention detection and mention clustering jointly and proposes a biaffine attention method to compute antecedent scores. Instead of using Word2vec and Elmo, Joshi et al. (2019, 2020) apply BERT to get better span representations.

There are also some efforts on solving the two tasks jointly. Iida and Poesio (2011) proposes an ILP-based model integrating the zero anaphora resolver with a coreference resolver. Kong and Ng (2013) exploits zero pronouns to improve non-zero coreference resolution by refining the syntactic parser and training examples. Shibata and Kurohashi (2018) presents an entity-based joint model for Japanese coreference resolution and predicate argument structure analysis. However, these works either rely on artificially designed features and syntactic information, or need the golden positions of the zero pronouns, which are costly and not practical. In this paper, we first propose an end-to-end neural model for solving the two tasks jointly.

3 Model

In this section, we first formally define the Zero & Non-Zero Joint Coreference Resolution task, and then present the details of the joint model.

3.1 Task Definition

This paper aims to solve the zero pronoun resolution and coreference resolution jointly. And we call the new task as **Zero & Non-Zero Joint Coreference Resolution**. Formally, given a doc-

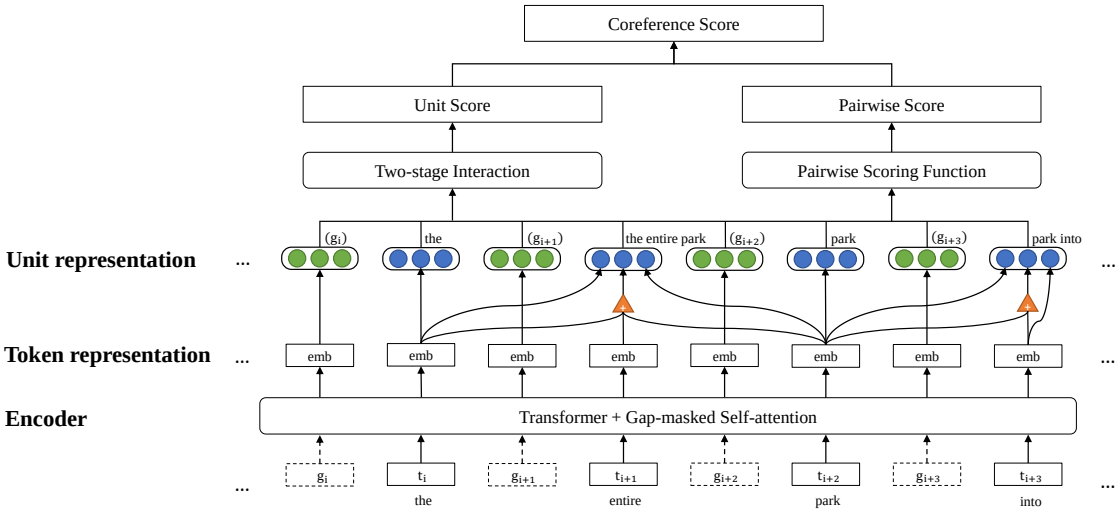

Figure 2: The architecture of end-to-end joint coreference resolution model

ument $D = \{t_1, t_2, t_3, ..., t_d\}$ where t_i represents the i-th token, $S = \{s_{11}, s_{12}, ..., s_{dd}\}$ is the set of spans, where s_{ij} means the span starting from the token t_i and ending at the token t_j and $G = \{g_1, g_2, ..., g_{d+1}\}$ is the set of gaps, where g_i means the gap before the token t_i and g_{d+1} is the last gap. For simplicity, we denote the set of spans and gaps as $U = \{u_1, u_2..., u_m\}$ where u_i is a unit which can be either a span or a gap.

Our task is to find a partition $P = \{X_1, X_2, ..., X_k, X_{k+1}\}$ from U, where k is the number of entities that appear in D. Each set X_i refers to an entity ent_i, and any two distinct sets X_i and X_j refer to different entities i.e. $ent_i \neq ent_j$. All units in the set $X_i (i \leq k)$ are the mentions or zero pronouns referring to the entity ent_i. There is a special entity called "empty entity" which the set X_{k+1} refers to, and the units in the set X_{k+1} are neither mentions nor zero pronouns.

3.2 End-to-end Joint Coreference Resolution

We introduce zero pronouns into the end-to-end neural coreference resolution model (Lee et al., 2018). Fig. 2 illustrates the framework of our model. The model first computes unit (span and gap) representations using transformer and gap-masked self-attention. Based on these embeddings, a unit score for each unit and a pairwise score for each pair of units will be calculated by a unit interaction mechanism and a pair-wise scoring function. These two scores are used to determine the antecedent.

The model learns the antecedent distribution $P(u_j)$ for each unit u_i:

$$P(u_j) = \frac{e^{s(u_i,u_j)}}{\sum_{u_j \in U_i} e^{s(u_i,u_j)}} \quad (1)$$

where U_i is the set of possible antecedent units for u_i which contains units in front of u_i and a dummy unit ϵ'. The pairwise score $s(u_i, u_j)$ represents the coreference score between u_i and u_j. The score is the sum of three factors:

$$s(u_i, u_j) = s_u(u_i) + s_u(u_j) + s_a(u_i, u_j) \quad (2)$$

where $s_u(u_i)$ is the unit score of u_i indicating the possibility that the unit u_i becomes mention or a zero pronoun and $s_a(u_i, u_j)$ is the pairwise score for u_i and u_j indicating the possibility that u_i and u_j refer to the same entity. Especially, the coreference score between dummy unit and any other unit is set to 0, i.e. $s(u_i, \epsilon') = 0$. In the basic joint model, these scoring functions are computed as follows:

$$s_u(u_i) = \begin{cases} FFNN_m(\boldsymbol{h}_{u_i}) & u_i \in S \\ FFNN_z(\boldsymbol{h}_{u_i}) & u_i \in G \end{cases}$$
$$s_a(u_i, u_j) = FFNN_a([\boldsymbol{h}_{u_i}; \boldsymbol{h}_{u_j}; \psi(u_i, u_j)]) \quad (3)$$

We use two different feed-forward neural networks $FFNN_m$ and $FFNN_z$ to score spans and gaps separately, where $\boldsymbol{h}_{u_i}$ is the span or gap representation for u_i and will be described in detail in the next section. While computing the pairwise score, we take the spans and gaps equally. The input of feed-forward neural network $FFNN_a$ consists of the

embeddings of two units u_i and u_j and the feature between u_i and u_j denoted as $\psi(u_i, u_j)$. Following Lee et al. (2018), a coarse-to-fine antecedent pruning approach and a representation refining approach are introduced to our model to reduce computational cost and enrich unit representation with global information.

3.3 Span & Gap Representations

As mentioned above, there are two kinds of units: span and gap. To reduce computational cost and prevent overfitting, we encode them into the same structure.

For computing a span u_i's representation, denoted as $\boldsymbol{h}_{u_i}$, we follow Lee et al. (2017):

$$\boldsymbol{h}_{u_i} = [\boldsymbol{e}_{START(u_i)}; \boldsymbol{e}_{END(u_i)}; \boldsymbol{e}_{ATT(u_i)}; \psi(u_i)] \tag{4}$$

where $\boldsymbol{e}_{START(u_i)}$ is the embedding of the first token for the span u_i and $\boldsymbol{e}_{END(u_i)}$ is the embedding of the last token for the span u_i. $\boldsymbol{e}_{ATT(u_i)}$ is an attention embedding computed over all tokens in span u_i. In addition to token embeddings, a feature vector $\psi(u_i)$ indicates the length of u_i is concatenated to the span representation.

For gap representations, we treat the gaps the same as the spans which only consist of a single token:

$$\boldsymbol{h}_{u_i} = [\boldsymbol{g}_{u_i}; \boldsymbol{g}_{u_i}; \boldsymbol{g}_{u_i}; \psi_{gap}] \tag{5}$$

where $\boldsymbol{g}_{u_i}$ is the gap embedding of u_i and has the same dimension as those of token embeddings. ψ_{gap} is a feature vector indicating that the unit is a gap (or the length of the unit is 0).

To get token embeddings $\boldsymbol{e}$ and gap embeddings $\boldsymbol{g}$, we firstly apply BERT to encode the document following the *independent* variant of splitting in Joshi et al. (2019) and get the basic token embeddings $\boldsymbol{e}'$. After that, we concatenate the token embedding $\boldsymbol{e}'_{i-1}$ with $\boldsymbol{e}'_i$, and map the concatenated embedding $[\boldsymbol{e}'_{i-1}; \boldsymbol{e}'_i]$ to the same dimension to get basic gap embedding $\boldsymbol{g}'_i$. Formally, the basic token embedding $\boldsymbol{e}'_i$ and the basic gap embedding $\boldsymbol{g}'_i$ are computed as follows:

$$\begin{aligned} \boldsymbol{e}'_1, \boldsymbol{e}'_2, ..., \boldsymbol{e}'_d &= BERT(t_1, t_2, ..., t_d) \\ \boldsymbol{g}'_i &= FFNN_g([\boldsymbol{e}'_{i-1}; \boldsymbol{e}'_i]) \end{aligned} \tag{6}$$

Based on basic embeddings $\boldsymbol{e}'$ and $\boldsymbol{g}'$, we use a gap-masked self-attention model to get final embeddings $\boldsymbol{e}$ and $\boldsymbol{g}$. The model is a variant of multi-head self-attention model where the weight for each gap is set to 0 while computing attention weights. That is, both token embeddings $\boldsymbol{e}$ and gap embeddings $\boldsymbol{g}$ are the weighted basic token embeddings. Therefore, gap embedding and token embedding can be in the same space without polluting token embedding. The details of the model are as follows:

$$\begin{aligned} \boldsymbol{e}, \boldsymbol{g} &= softmax(\boldsymbol{W})\boldsymbol{V} \\ \boldsymbol{W}_i &= \begin{cases} \dfrac{\boldsymbol{Q}\boldsymbol{K}_i^T}{\sqrt{d_k}} & i \in D \\ -\inf & i \in G \end{cases} \\ \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} &= linear_{q,k,v}(\{\boldsymbol{e}', \boldsymbol{g}'\}) \end{aligned} \tag{7}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are $query, key, value$ following Vaswani et al. (2017), and all tokens and gaps will get these three factors based on $\boldsymbol{e}'$ and $\boldsymbol{g}'$ to calculate attention weights $\boldsymbol{W}$. $linear_{q,k,v}$ are three different linear functions to compute $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$. $\sqrt{d_k}$ is the scaling factor and d_k is the size of $\boldsymbol{K}$.

3.4 Two-stage Interaction Mechanism

As mentioned in Section 1, there is an exclusive relationship between zero pronouns and mentions. Therefore, the unit score of a span should be affected by the gaps in it, and the unit score of a gap will be affected by the spans across it. We define these gaps and spans as the relevant unit set R_{u_i} for the unit u_i:

$$R_{u_i} = \begin{cases} \{g_x | g_x \in G, p < x <= q\} & u_i = s_{pq} \\ \{s_{xy} | s_{xy} \in S, x < p <= y\} & u_i = g_p \end{cases} \tag{8}$$

When the unit u_i is a span consisting of a single token ($s_{pq}, p = q$) or the gap at the beginning or the end of the document (g_0, g_{d+1}), the relevant units set R_{u_i} will be an empty set.

To make use of the exclusive relationship between zero pronouns and mentions, we propose a two-stage interaction mechanism. By modifying the unit scoring function s_u, the mechanism will introduce the relevant units of u_i when computing the unit score $s_u(u_i)$, so that the unit score of u_i can be affected by its relevant units. According to the incorporation of relevant units, the interaction mechanism includes two different interaction methods: the interaction among unit representations and the interaction among unit scores. For the interaction among unit representations, the representations of relevant units will be used when the model computes unit scores. We concatenate the unit embedding with the attention embedding of

its relevant units and take the concatenated result as the input of the scoring function. If the relevant unit set is an empty set, we then use a special embedding ϵ instead of an attention embedding. Therefore, the unit scoring function s_u is modified as s'_u:

$$
\begin{aligned}
s'_u(u_i) &= \begin{cases} FFNN_m([\boldsymbol{h}_{u_i}; \boldsymbol{h}_{R_{u_i}}]) & u_i \in S \\ FFNN_z([\boldsymbol{h}_{u_i}; \boldsymbol{h}_{R_{u_i}}]) & u_i \in G \end{cases} \\
\boldsymbol{h}_{R_{u_i}} &= \sum_{u_j \in R_{u_i}} softmax(\alpha_{u_j}) \boldsymbol{h}_{u_j} \\
\alpha_{u_j} &= FFNN_\alpha(\boldsymbol{h}_{u_j})
\end{aligned} \tag{9}
$$

For the interaction among unit scores, the unit scores of relevant units will be used when the model computes the unit score. We update the unit score of u_i by the score over its relevant units R_{u_i}. Intuitively, there is a negative correlation between these two unit scores. The higher score of a unit brings the lower score over its relevant units:

$$
s''_u(u_i) = s'_u(u_i) - I(\bigcup_{u_j \in R_{u_i}} s'_u(u_j)) \tag{10}
$$

where I is an aggregate function with unit scores of relevant units as input. In our model, we adopt two common functions to aggregate these unit scores: max and $mean$. The max function outputs the highest unit score, and the $mean$ function calculates average unit scores.

3.5 Training

Our model trains resolution for zero pronouns and mentions simultaneously. However, the number of zero pronouns is far less than mentions in the training data. If we only adopt the clustering loss (Lee et al., 2017) which considers antecedent links, the learning process of zero pronoun resolution will be slow. Therefore, we follow Zhang et al. (2018), combining unit detection loss with clustering loss:

$$
L = \lambda \sum_{i \in U} L_{detect}(i) + \sum_{i \in U'} L_{cluster}(i) \tag{11}
$$

where λ controls the weights of two loss, U is the set of all units, and U' is the set of units after coarse-to-fine pruning (Lee et al., 2018).

For the detection loss L_{detect}:

$$
L_{detect}(i) = -y_i log(\hat{y}_i) - (1 - y_i) log(1 - \hat{y}_i) \tag{12}
$$

where y_i is the gold label of unit i, $y_i = 1$ means that unit i is a zero pronoun or a mention, otherwise $y_i = 0$. $\hat{y}_i = sigmoid(s_u(i))$ is the predicted unit score of unit i.

For the clustering loss $L_{cluster}$:

$$
L_{cluster}(i) = -log \sum_{u \in U_i \cap GOLD(i)} P(u) \tag{13}
$$

where U_i is the set of possible antecedent units for i, and $GOLD(i)$ is the set of gold antecedent units for i. If i is either a singleton mention or a singleton zero pronoun, or i is neither a mention nor a zero pronoun, or gold antecedents are all pruned, $GOLD(i) = \epsilon'$.

3.6 Evaluating & Predicting

There are two different antecedents setting strategies, "To ZP" and "Not to ZP", according to whether the zero pronouns are allow to be resolved to as antecedents. The "To ZP" strategy stipulates that zero pronouns can be resolved to if they get the top coreference scores while the "Not to ZP" strategy stipulates that zero pronouns cannot be resolved to even if they get the top coreference scores. For example in Fig. 1, to find the best antecedent of m_5, the model calculates all coreference scores for m_5: $\bigcup_{u \in U_{m_5}} s(m_5, u)$, where $U_{m_5} = \{m_1, m_2, m_3, m_4, \phi_1, \phi_2\}$. We assume that $s(m_5, \phi_2) = 0.6$, $s(m_5, m_3) = 0.3$, and we ignore the other scores. If we adopt the "To ZP" strategy, the model would choose zero pronoun ϕ_2 as the best antecedent. But if we choose the "Not to ZP" strategy, the model would skip all zero pronouns and choose the mention with the highest score as the best antecedent (in this example the best antecedent is m_3).

4 Experiments

In this section, we study the effectiveness of our model. We firstly introduce the datasets and the metrics we use to evaluate models. Then we present the details of experimental settings in the model. Finally, we display our experimental results.

4.1 Datasets

We train and evaluate our model on the OntoNotes 5.0 Chinese corpus[1]. The corpus is split into train/dev/test datasets[2] which contain 1810 training documents, 250 development documents, and 218 testing documents.

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

[2] We use same algorithm as in CoNLL-2012 to create training, development and test partitions

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
(Clark and Manning, 2016b)	73.85	65.42	69.38	67.53	56.41	61.47	62.84	57.62	60.12	63.66
(Clark and Manning, 2016a)	73.64	65.62	69.40	67.48	56.94	61.76	62.46	58.60	60.47	63.88
(Kong and Fu, 2019)	76.95	64.58	70.21	**70.58**	54.68	61.60	64.92	55.36	59.75	63.85
(Joshi et al., 2019)	76.33	65.53	70.52	67.15	55.99	61.07	65.65	53.47	58.94	63.51
+ hyperparameter tuning	74.42	**70.80**	72.57	65.68	62.74	64.18	66.98	59.61	63.08	66.61
Our model	77.07	68.83	72.71	68.79	61.28	64.82	69.62	60.38	64.67	67.40
+ IUR	75.33	69.76	72.44	68.17	62.30	65.10	68.69	**62.20**	**65.29**	67.61
+ IUS(max)	75.53	70.55	**72.96**	66.79	**63.28**	64.99	68.93	61.81	65.17	**67.71**
+ IUS(mean)	**77.77**	68.31	72.74	70.40	60.85	**65.28**	**69.58**	61.15	65.09	67.70

Table 1: Results of non-zero coreference resolution on Chinese corpus in OntoNotes 5.0. The average F1 of MUC, B^3, and $CEAF_{\phi_4}$ is the main evaluation metric.

	P	R	F1
(Yang et al., 2019)	-	-	17.25
(Song et al., 2020)	30.96	22.51	26.07
Our baseline	36.22	30.84	33.31
Our model	37.00	31.06	33.77
+ IUR	36.74	**31.63**	33.99
+ IUS(max)	**37.56**	31.28	**34.13**
+ IUS(mean)	37.51	31.25	34.10

Table 2: Results of zero pronoun resolution on Chinese corpus in OntoNotes 5.0.

	M-F1	ZP-F1
Our model	78.78	47.73
+ IUR	78.83	47.90
+ IUS(max)	**78.86**	47.99
+ IUS(mean)	78.84	**48.00**

Table 3: Results of mention detection and zero pronoun detection, where "M-F1" represents F1 score of mention detection and "ZP-F1" represents F1 score of zero pronoun detection.

4.2 Evaluation Metrics

The metrics for evaluating mentions and zero pronouns are different.

To evaluate coreference results of mentions, we follow the previous coreference resolution works including three metrics: MUC, B^3, and $CEAF_{\phi_4}$. We report the precision, recall and F-score for each metric. Also, we report the average F-score of the three metrics.

To evaluate zero pronoun resolution results, we follow the previous zero pronoun resolution works (Zhao and Ng, 2007; Chen and Ng, 2013) to report precision, recall and F-score.

In addition, we also report precision, recall and F-score for zero pronoun detection and mention detection.

4.3 Experimental Settings

Our model reuses most of the hyperparameters from Joshi et al. (2019) except: (1) We use the official pretrained Chinese BERT-base model[3] to encode tokens. (2) To reduce computational cost and memory usage, we decrease the maximum span width from 30 to 20 tokens. (3) Due to the introduction of zero pronouns, we increase the top unit ratio, which indicates the ratio of units being kept after pruning, from 0.4 to 0.5. (4) The gap-masked self-attention model uses 1 self-attention layer with 8 heads, and is trained with AdamW (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-4}. (5) The weight of detection loss λ is set to 0.2.

4.4 Main Results

The main results on test sets are shown in Table 1 and Table 2. In particular, Table 1 shows the results of non-zero coreference resolution. We compare our model with previous works which are trained and tested only on non-zero part of the OntoNotes 5.0 Chinese datasets. The baseline is an end-to-end neural coreference resolution model (Joshi et al., 2019) using official Chinese BERT-base model. For fair comparison, we tune hyperparameters and add a self-attention model which has the same size with the gap-masked self-attention model after the BERT encoding layer on the baseline model. Compared with the baseline model which only resolves mentions, the joint model has better performance on non-zero coreference resolution. Benefited from the introduction of zero pronouns, the joint model outperforms the baseline by at least 0.8%.

Table 2 shows the results of zero pronoun resolu-

[3]https://github.com/google-research/bert

	NZ-F1	Z-F1
Gaps embedded BERT	58.96	27.96
FFNN + tanh	64.09	31.07
LSTM	64.15	31.21
Self-attention	66.42	32.15
Gap-masked self-attention	**67.40**	**33.77**

Table 4: Results of different encoders. NZ-F1 and Z-F1 represent the F1 on non-zero coreference resolution and zero pronoun resolution respectively

	To ZP		Not to ZP	
	NZ-F1	Z-F1	NZ-F1	Z-F1
Our model	67.38	33.71	67.40	33.77
+ IUR	67.51	33.59	67.61	33.99
+ IUS(max)	67.65	34.12	67.71	34.13
+ IUS(mean)	67.68	34.03	67.70	34.10

Table 5: Results of coreference resolution and zero pronoun resolution with different antecedents setting strategies

tion, where Song et al. (2020) proposes the state-of-the-art model in zero pronoun resolution without relying on syntactic information. We remove the training process of non-zero coreference resolution from our model as our baseline. The baseline joint model greatly improves the performance of zero pronoun resolution, which demonstrates the advantages of having mentions detected for the task in joint learning. The results of the joint model also demonstrate the positive effects of non-zero coreference resolution on zero pronoun resolution, which could improve the zero pronoun resolution by more than 0.46%.

We can also observe that the introduction of two-stage interaction mechanism produces better performance. In Table 1, Table 2 and Table 5, "+ IUR" means using the interaction among unit representations described in Eq. 9, and "+ IUS" means using the interaction among unit scores described in Eq. 10. We test two different aggregate functions max and $mean$ while using the interaction among unit scores. According to the results, we can see that both of the aggregate functions max and $mean$ can improve the joint model, and using the max aggregate function has a better performance than using the $mean$ aggregate function. Table 3 shows that the introduction of two-stage interaction mechanism improves both mention detection and zero pronoun detection results.

4.5 Encoder Evaluation

As shown in Table 4, we apply different encoders to get representations for tokens and gaps in our model. "Gaps embedded BERT" encoder inserts all gaps into documents directly and uses one BERT model to encode gaps and tokens simultaneously. The weights for gaps are non-zero unlike the gap-masked self-attention encoder used in our model. With this encoder, embeddings of gaps and tokens can be encoded into the same space but the insertion of gaps will pollute the representations of tokens. "FFNN + tanh" encoder uses feed-forward neural networks and $tanh$ activation function to map token embeddings into gap embeddings. "LSTM" encoder and "Self-attention" encoder apply LSTM or self-attention model on token embeddings to get gap embeddings. These three encoders compute gap embeddings without updating token embeddings obtained by BERT, where token representations will not be polluted but embeddings of gaps and tokens will be in different space. "Gap-masked self-attention" is the encoder used in our model. It can encode gaps and update token representations at the same time without polluting token representations and can put embeddings into the same space.

4.6 Antecedents Setting Strategy Analysis

As mentioned in Section 3.6, we select different antecedents setting strategies in the evaluation. Table 5 shows the results of adopting different strategies. As we can see, the "Not to ZP" strategy has better performance in all tasks and models. The reason is that the performance of zero pronoun detection and zero pronoun resolution is poor compared with that of mention detection and non-zero coreference resolution. Therefore, resolving to zero pronouns will introduce more errors.

5 Conclusions

In this paper, we present an end-to-end joint coreference resolution model which tackles zero pronoun resolution and non-zero coreference resolution jointly. To get proper representations for gaps and tokens, we propose a gap-masked self-attention model which puts embeddings of tokens and gaps into the same space without polluting token representations. Additionally, to make full use of the exclusive relationship between zero pronoun and mentions, we propose a two-stage interaction mechanism which incorporates information of relevant

units while calculating unit scores. Extensive experiments on OntoNotes 5.0 Chinese corpus demonstrate the effectiveness of our model.

Acknowledgments

This work was supported by the National Key R&D Program of China (No.2018AAA0101900), the Priority Academic Program Development of Jiangsu Higher Education Institutions, National Natural Science Foundation of China (Grant No. 62072323, 61632016, 62102276), Natural Science Foundation of Jiangsu Province (No. BK20191420).

References

Abdulrahman Aloraini and Massimo Poesio. 2020. Cross-lingual zero pronoun resolution. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 90–98, Marseille, France. European Language Resources Association.

Saoussen Mathlouthi Bouzid and Chiraz Ben Othmane Zribi. 2020. A generic approach for pronominal anaphora and zero anaphora resolution in arabic language. Procedia Computer Science, 176:642–652.

Chen Chen and Vincent Ng. 2013. Chinese zero pronoun resolution: Some recent advances. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1360–1365, Seattle, Washington, USA. Association for Computational Linguistics.

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 778–788, Berlin, Germany. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 643–653, Berlin, Germany. Association for Computational Linguistics.

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, pages 804–813.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Fang Kong and Jian Fu. 2019. Incorporating structural information for better coreference resolution. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pages 5039–5045. ijcai.org.

Fang Kong and Hwee Tou Ng. 2013. Exploiting zero pronouns to improve Chinese coreference resolution. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 278–288, Seattle, Washington, USA. 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, Cambridge, MA. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Peiqin Lin and Meng Yang. 2020. Hierarchical attention network with pairwise loss for chinese zero pronoun resolution. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8352–8359. AAAI Press.

Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 102–111, Vancouver, Canada. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Ruslan Mitkov et al. 1995. Anaphora resolution in machine translation. In Proceedings of the Sixth International conference on Theoretical and Methodological issues in Machine Translation. Citeseer.

Tomohide Shibata and Sadao Kurohashi. 2018. Entity-centric joint modeling of japanese coreference resolution and predicate argument structure analysis. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 579–589.

Linfeng Song, Kun Xu, Yue Zhang, Jianshu Chen, and Dong Yu. 2020. ZPR2: Joint zero pronoun recovery and resolution using multi-task learning and BERT. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5429–5434, Online. Association for Computational Linguistics.

Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek. 2007. Two uses of anaphora resolution in summarization. Information Processing & Management, 43(6):1663–1680.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. Information Fusion, 59:139–162.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2019. Coreference resolution as query-based span prediction. arXiv preprint arXiv:1911.01746.

Wei Yang, Rui Qiao, Haocheng Qin, Amy Sun, Luchen Tan, Kun Xiong, and Ming Li. 2019. End-to-end neural context reconstruction in Chinese dialogue. In Proceedings of the First Workshop on NLP for Conversational AI, pages 68–76, Florence, Italy. Association for Computational Linguistics.

Qingyu Yin, Weinan Zhang, Yu Zhang, and Ting Liu. 2017. A deep neural network for chinese zero pronoun resolution. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pages 3322–3328. ijcai.org.

Qingyu Yin, Yu Zhang, Wei-Nan Zhang, Ting Liu, and William Yang Wang. 2018a. Deep reinforcement learning for Chinese zero pronoun resolution. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 569–578, Melbourne, Australia. Association for Computational Linguistics.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018b. Zero pronoun resolution with attention-based neural network. In Proceedings of the 27th International Conference on Computational Linguistics, pages 13–23, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. 2004. Coreference resolution for information extraction. In Proceedings of the Conference on Reference Resolution and Its Applications, pages 24–31, Barcelona, Spain. Association for Computational Linguistics.

Rui Zhang, Cícero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. 2018. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 102–107, Melbourne, Australia. Association for Computational Linguistics.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of chinese zero pronouns: A machine learning approach. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pages 541–550.