

# Incorporating Context and External Knowledge for Pronoun Coreference Resolution

Hongming Zhang<sup>♣\*</sup>, Yan Song<sup>♣</sup>, and Yangqiu Song<sup>♣</sup>

<sup>♣</sup>Department of CSE, The Hong Kong University of Science and Technology

<sup>♣</sup>Tencent AI Lab

hzhangal@cse.ust.hk, clksong@gmail.com, yqsong@cse.ust.hk

## Abstract

Linking pronominal expressions to the correct references requires, in many cases, better analysis of the contextual information and external knowledge. In this paper, we propose a two-layer model for pronoun coreference resolution that leverages both context and external knowledge, where a knowledge attention mechanism is designed to ensure the model leveraging the appropriate source of external knowledge based on different context. Experimental results demonstrate the validity and effectiveness of our model, where it outperforms state-of-the-art models by a large margin.

## 1 Introduction

The question of how human beings resolve pronouns has long been of interest to both linguistics and natural language processing (NLP) communities, for the reason that pronoun itself has weak semantic meaning (Ehrlich, 1981) and brings challenges in natural language understanding. To explore solutions for that question, pronoun coreference resolution (Hobbs, 1978) was proposed. As an important yet vital sub-task of the general coreference resolution task, pronoun coreference resolution is to find the correct reference for a given pronominal anaphor in the context and has been shown to be crucial for a series of downstream tasks (Mitkov, 2014), including machine translation (Mitkov et al., 1995), summarization (Steinberger et al., 2007), information extraction (Edens et al., 2003), and dialog systems (Strube and Müller, 2003).

Conventionally, people design rules (Hobbs, 1978; Nasukawa, 1994; Mitkov, 1998) or use features (Ng, 2005; Charniak and Elsner, 2009; Li et al., 2011) to resolve the pronoun coreferences.

\*This work was done during the internship of the first author in Tencent AI Lab.

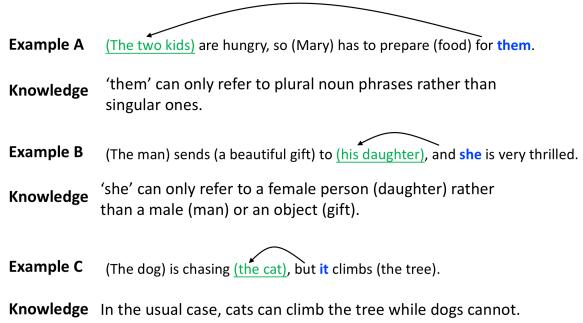


Figure 1: Pronoun coreference examples, where each example requires different knowledge for its resolution. Blue bold font refers to the target pronoun, where the correct noun reference and other candidates are marked by green underline and brackets, respectively.

These methods heavily rely on the coverage and quality of the manually defined rules and features. Until recently, end-to-end solution (Lee et al., 2017) was proposed towards solving the general coreference problem, where deep learning models were used to better capture contextual information. However, training such models on annotated corpora can be biased and normally does not consider external knowledge.

Despite the great efforts made in this area in the past few decades (Hobbs, 1978; Mitkov, 1998; Ng, 2005; Rahman and Ng, 2009), pronoun coreference resolution remains challenging. The reason behind is that the correct resolution of pronouns can be influenced by many factors (Ehrlich, 1981); many resolution decisions require reasoning upon different contextual and external knowledge (Rahman and Ng, 2011), which is also proved in other NLP tasks (Song et al., 2017, 2018; Zhang et al., 2018). Figure 1 demonstrates such requirement with three examples, where Example A depends on the plurality knowledge that ‘them’ refers to plural noun phrases; Example B illustrates the gender requirement of pronouns where ‘she’ can

only refer to a female person (girl); Example C requires a more general type of knowledge<sup>1</sup> that ‘cats can climb trees but a dog normally does not’. All of these knowledge are difficult to be learned from training data. Considering the importance of both contextual information and external human knowledge, how to jointly leverage them becomes an important question for pronoun coreference resolution.

In this paper, we propose a two-layer model to address the question while solving two challenges of incorporating external knowledge into deep models for pronoun coreference resolution, where the challenges include: first, different cases have their knowledge preference, i.e., some knowledge is exclusively important for certain cases, which requires the model to be flexible in selecting appropriate knowledge per case; second, the availability of knowledge resources is limited and such resources normally contain noise, which requires the model to be robust in learning from them.

Consequently, in our model, the first layer predicts the relations between candidate noun phrases and the target pronoun based on the contextual information learned by neural networks. The second layer compares the candidates pair-wisely, in which we propose a knowledge attention module to focus on appropriate knowledge based on the given context. Moreover, a softmax pruning is placed in between the two layers to select high confident candidates. The architecture ensures the model being able to leverage both context and external knowledge. Especially, compared with conventional approaches that simply treat external knowledge as rules or features, our model is not only more flexible and effective but also interpretable as it reflects which knowledge source has the higher weight in order to make the decision. Experiments are conducted on a widely used evaluation dataset, where the results prove that the proposed model outperforms all baseline models by a great margin.<sup>2</sup>

Above all, to summarize, this paper makes the following contributions:

1. We propose a two-layer neural model to combine contextual information and external

<sup>1</sup>This is normally as selectional preference (SP) (Hobbs, 1978), which is defined as given a predicate (verb), a human has the preference for its argument (subject in this example).

<sup>2</sup>All code and data are available at:<https://github.com/HKUST-KnowComp/Pronoun-Coref>.

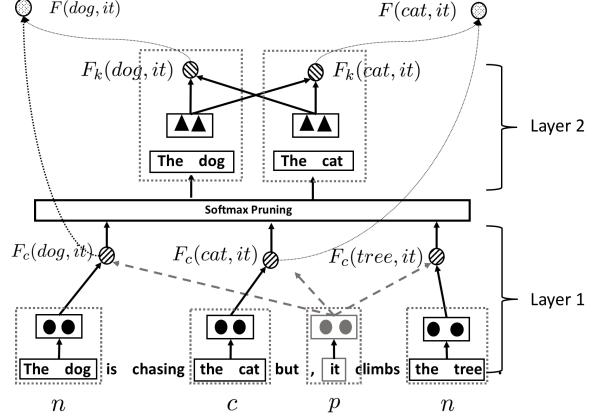


Figure 2: The architecture of the two-layer model for pronoun coreference resolution. The first layer encodes the contextual information for computing  $F_c$ . The second layer leverages external knowledge to score  $F_k$ . A pruning layer is applied in between the two layers to control computational complexity. The dashed boxes in the first and second layer refer to span representation and knowledge scoring, respectively.

knowledge for the pronoun coreference resolution task.

2. We propose a knowledge attention mechanism that allows the model to select salient knowledge for different context, which predicts more precisely and can be interpretable through the learned attention scores.
3. With our proposed model, the performance of pronoun coreference resolution is boosted by a great margin over the state-of-the-art models.

## 2 The Task

Following the conventional setting (Hobbs, 1978), the task of pronoun coreference resolution is defined as: for a pronoun  $p$  and a candidate noun phrase set  $\mathcal{N}$ , the goal is to identify the correct non-pronominal references set<sup>3</sup>  $\mathcal{C}$ . the objective is to maximize the following objective function:

$$\mathcal{J} = \frac{\sum_{c \in \mathcal{C}} e^{F(c,p)}}{\sum_{n \in \mathcal{N}} e^{F(n,p)}}, \quad (1)$$

where  $c$  is the correct reference and  $n$  the candidate noun phrase.  $F(\cdot)$  refers to the overall coreference scoring function for each  $n$  regarding  $p$ . Following (Mitkov, 1998), all non-pronominal noun phrases in the recent three sentences of the pronoun  $p$  are selected to form  $\mathcal{N}$ .

Particularly in our setting, we want to leverage both the local contextual information and external

<sup>3</sup>It is possible that a pronoun has multiple references.

knowledge in this task, thus for each  $n$  and  $p$ ,  $F(\cdot)$  is decomposed into two components:

$$F(n, p) = F_c(n, p) + F_k(n, p), \quad (2)$$

where  $F_c(n, p)$  is the scoring function that predicts the relation between  $n$  and  $p$  based on the contextual information;  $F_k(n, p)$  is the scoring function that predicts the relation between  $n$  and  $p$  based on the external knowledge. There could be multiple ways to compute  $F_c$  and  $F_k$ , where a solution proposed in this paper is described as follows.

### 3 The Model

The architecture of our model is shown in Figure 2, where we use two layers to incorporate contextual information and external knowledge. Specifically, the first layer takes the representations of different  $n$  and the  $p$  as input and predict the relationship between each pair of  $n$  and  $p$ , so as to compute  $F_c$ . The second layer leverages the external knowledge to compute  $F_k$ , which consists of pair-wise knowledge score  $f_k$  among all candidate  $n$ . To enhance the efficiency of the model, a softmax pruning module is applied to select high confident candidates into the second layer. The details of the aforementioned components are described in the following subsections.

#### 3.1 Encoding Contextual Information

Before  $F_c$  is computed, the contextual information is encoded through a span<sup>4</sup> representation (SR) module in the first layer of the model. Following Lee et al. (2017), we adopt the standard bidirectional LSTM (biLSTM) (Hochreiter and Schmidhuber, 1997) and the attention mechanism (Bahdanau et al., 2015) to generate the span representation, as shown in Figure 3. Given that the initial word representations in a span  $n_i$  are  $\mathbf{x}_1, \dots, \mathbf{x}_T$ ,

we denote their representations  $\mathbf{x}_1^*, \dots, \mathbf{x}_T^*$  after encoded by the biLSTM. Then we obtain the inner-span attention by

$$a_t = \frac{e^{\alpha_t}}{\sum_{k=1}^T e^{\alpha_k}}, \quad (3)$$

where  $\alpha_t$  is computed via a standard feed-forward neural network<sup>5</sup>  $\alpha_t = NN_\alpha(\mathbf{x}_t^*)$ . Thus, we have

<sup>4</sup>Both noun phrases and the pronoun are treated as spans.

<sup>5</sup>We use  $NN$  to present feed-forward neural networks throughout this paper.

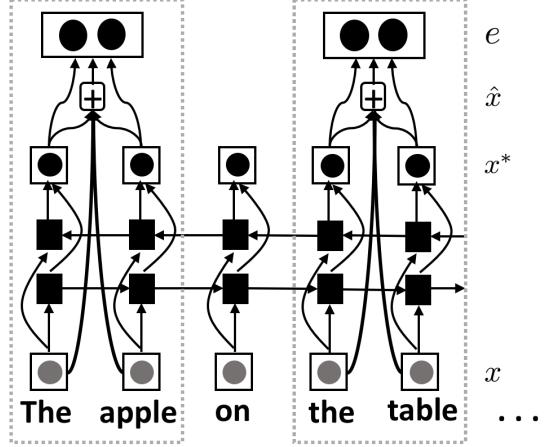


Figure 3: The structure of span representation. Bidirectional LSTM and inner-span attention mechanism are employed to capture the contextual information.

the weighted embedding of each span  $\hat{x}_i$  through

$$\hat{\mathbf{x}}_i = \sum_{k=1}^T a_k \cdot \mathbf{x}_k. \quad (4)$$

Afterwards, we concatenate the starting ( $\mathbf{x}_{start}^*$ ) and ending ( $\mathbf{x}_{end}^*$ ) embedding of each span, as well as its weighted embedding ( $\hat{\mathbf{x}}_i$ ) and the length feature ( $\phi(i)$ ) to form its final representation  $e$ :

$$\mathbf{e}_i = [\mathbf{x}_{start}^*, \mathbf{x}_{end}^*, \hat{\mathbf{x}}_i, \phi(i)]. \quad (5)$$

Once the span representation of  $n \in \mathcal{N}$  and  $p$  are obtained, we compute  $F_c$  for each  $n$  with a standard feed-forward neural network:

$$F_c(n, p) = NN_c([\mathbf{e}_n, \mathbf{e}_p, \mathbf{e}_n \odot \mathbf{e}_p]), \quad (6)$$

where  $\odot$  is the element-wise multiplication.

#### 3.2 Processing External Knowledge

In the second layer of our model, external knowledge is leveraged to evaluate all candidate  $n$  so as to give them reasonable  $F_k$  scores. In doing so, each candidate is represented as a group of features from different knowledge sources, e.g., ‘the cat’ can be represented as a singular noun, unknown gender creature, and a regular subject of the predicate verb ‘climb’. For each candidate, we conduct a series of pair-wise comparisons between it and all other ones to result in its  $F_k$  score. An attention mechanism is proposed to perform the comparison and selectively use the knowledge features. Consider there exists noise in external knowledge, especially when it is automatically

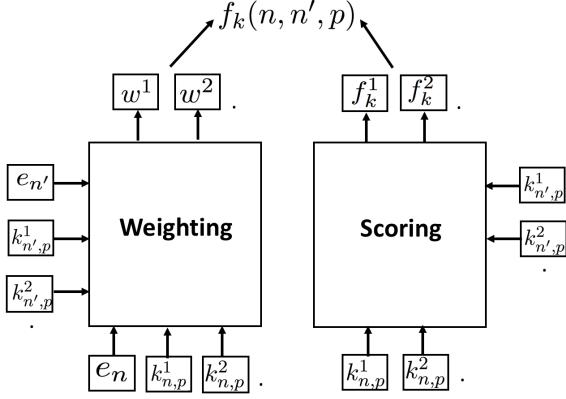


Figure 4: The structure of the knowledge attention module. For each feature  $k_i$  from knowledge source  $i$ , the weighting component predict its weight  $w^i$  and the scoring component computes its knowledge score  $f_k^i$ . Then a weighted sum is obtained for  $f_k$ .

generated, such attention mechanism ensures that, for each candidate, reliable and useful knowledge is utilized rather than ineffective ones. The details of the knowledge attention module and the overall scoring are described as follows.

**Knowledge Attention** Figure 4 demonstrates the structure of the knowledge attention module, where there are two components: (1) weighting: assigning weights to different knowledge features regarding their importance in the comparison; (2) scoring: valuing a candidate against another one based on their features from different knowledge sources. Assuming that there are  $m$  knowledge sources input to our model, each candidate can be represented by  $m$  different features, which are encoded as embeddings. Therefore, two candidates  $n$  and  $n'$  regarding  $p$  have their knowledge feature embeddings  $\mathbf{k}_{n,p}^1, \mathbf{k}_{n,p}^2, \dots, \mathbf{k}_{n,p}^m$  and  $\mathbf{k}_{n',p}^1, \mathbf{k}_{n',p}^2, \dots, \mathbf{k}_{n',p}^m$ , respectively. The weighting component receives all features  $\mathbf{k}$  for  $n$  and  $n'$ , and the span representations  $\mathbf{e}_n$  and  $\mathbf{e}_{n'}$  as input, where  $\mathbf{e}_n$  and  $\mathbf{e}_{n'}$  help selecting appropriate knowledge based on the context. As a result, for a candidate pair  $(n, n')$  and a knowledge source  $i$ , its knowledge attention score is computed via

$$\beta_i(n, n', p) = NN_{ka}([\mathbf{o}_{n,p}^i, \mathbf{o}_{n',p}^i, \mathbf{o}_{n,p}^i \odot \mathbf{o}_{n',p}^i]), \quad (7)$$

where  $\mathbf{o}_{n,p}^i = [\mathbf{e}_n, \mathbf{k}_{n,p}^i]$  and  $\mathbf{o}_{n',p}^i = [\mathbf{e}_{n'}, \mathbf{k}_{n',p}^i]$  are the concatenation of span representation and external knowledge embedding for candidate  $n$  and  $n'$  respectively. The weight for features from

different knowledge sources is thus computed via

$$w_i = \frac{e^{\beta_i}}{\sum_{j=1}^m e^{\beta_j}}. \quad (8)$$

Similar to the weighting component, for each feature  $i$ , we compute its score  $f_k^i(n, n', p)$  for  $n$  against  $n'$  in the scoring component through

$$f_k^i(n, n', p) = NN_{ks}([\mathbf{k}_{n,p}^i, \mathbf{k}_{n',p}^i, \mathbf{k}_{n,p}^i \odot \mathbf{k}_{n',p}^i]). \quad (9)$$

where it is worth noting that we exclude  $\mathbf{e}$  in this component for the reason that, in practice, the dimension of  $\mathbf{e}$  is normally much higher than  $\mathbf{k}$ . As a result, it could dominate the computation if  $\mathbf{e}$  and  $\mathbf{k}$  is concatenated.<sup>6</sup>

Once the weights and scores are obtained, we have a weighted knowledge score for  $n$  against  $n'$ :

$$f_k(n, n', p) = \sum_{i=1}^m w_i \cdot f_k^i(n, n', p). \quad (10)$$

**Overall Knowledge Score** After all pairs of  $n$  and  $n'$  are processed by the attention module, the overall knowledge score for  $n$  is computed through the averaged  $f_k(n, n', p)$  over all  $n'$ :

$$F_k(n, p) = \frac{\sum_{n' \in \mathcal{N}_o} f_k(n, n', p)}{|\mathcal{N}_o|}, \quad (11)$$

where  $\mathcal{N}_o = \mathcal{N} - n$  for each  $n$ .

### 3.3 Softmax Pruning

Normally, there could be many noun phrases that serve as the candidates for the target pronoun. One potential obstacle in the pair-wise comparison of candidate noun phrases in our model is the squared complexity  $O(|\mathcal{N}|^2)$  with respect to the size of  $\mathcal{N}$ . To filter out low confident candidates so as to make the model more efficient, we use a softmax-pruning module between the two layers in our model to select candidates for the next step. The module takes  $F_c$  as input for each  $n$ , uses a softmax computation:

$$\hat{F}_c(n, p) = \frac{e^{F_c(n, p)}}{\sum_{n_i \in \mathcal{N}} e^{F_c(n_i, p)}}. \quad (12)$$

where candidates with higher  $\hat{F}_c$  are kept, based on a threshold  $t$  predefined as the pruning standard. Therefore, if candidates have similar  $F_c$

<sup>6</sup>We do not have this concern for the weighting component because the softmax (c.f. Eq. 8) actually amplifies the difference of  $\beta$  even if they are not much differentiated.

type	train	dev	test	all
Third Personal Possessive	21,828 7,749	2,518 1,007	3,530 1,037	27,876 9,793
All	29,577	3,525	4,567	37,669

Table 1: Statistics of the evaluation dataset. Number of selected pronouns are reported.

scores, the module allow more of them to proceed to the second layer. Compared with other conventional pruning methods (Lee et al., 2017, 2018) that generally keep a fixed number of candidates, our pruning strategy is more efficient and flexible.

## 4 Experiment Settings

### 4.1 Data

The CoNLL-2012 shared task (Pradhan et al., 2012) corpus is used as the evaluation dataset, which is selected from the Ontonotes 5.0<sup>7</sup>. Following conventional approaches (Ng, 2005; Li et al., 2011), for each pronoun in the document, we consider candidate  $n$  from the previous two sentences and the current sentence. For pronouns, we consider two types of them following Ng (2005), i.e., third personal pronoun (*she, her, he, him, them, they, it*) and possessive pronoun (*his, hers, its, their, theirs*). Table 1 reports the number of the two type pronouns and the overall statistics for the experimental dataset. According to our selection range of candidate  $n$ , on average, each pronoun has 4.6 candidates and 1.3 correct references.

### 4.2 Knowledge Types

In this study, we use two types of knowledge in our experiments. The first type is linguistic features, i.e., plurality and animacy & gender. We employ the Stanford parser<sup>8</sup>, which generates plurality, animacy, and gender markups for all the noun phrases, to annotate our data. Specifically, the plurality feature denotes each  $n$  and  $p$  to be singular or plural. For each candidate  $n$ , if its plurality status is the same as the target pronoun, we label it 1, otherwise 0. The animacy & gender (AG) feature denotes whether a  $n$  or  $p$  is a living object, and being male, female, or neutral if it is alive. For each candidate  $n$ , if its AG feature matches the target pronoun’s, we label it 1, otherwise 0.

The second type is the selectional preference (SP) knowledge. For this knowledge, we create

<sup>7</sup><https://catalog.ldc.upenn.edu/LDC2013T19>

<sup>8</sup><https://stanfordnlp.github.io/CoreNLP/>

a knowledge base by counting how many times a predicate-argument tuple appears in a corpus and use the resulted number to represent the preference strength. Specifically, we use the English Wikipedia<sup>9</sup> as the base corpus for such counting. Then we parse the entire corpus through the Stanford parser and record all dependency edges in the format of  $(\text{predicate}, \text{argument}, \text{relation}, \text{number})$ , where predicate is the governor and argument the dependent in the original parsed dependency edge<sup>10</sup>. Later for sentences in the training and test data, we firstly parse each sentence and find out the dependency edge linking  $p$  and its corresponding predicate. Then for each candidate<sup>11</sup>  $n$  in a sentence, we check the previously created SP knowledge base and find out how many times it appears as the argument of different predicates with the same dependency relation (i.e., *nsubj* and *dobj*). The resulted frequency is grouped into the following buckets [1, 2, 3, 4, 5-7, 8-15, 16-31, 32-63, 64+] and we use the bucket id as the final SP knowledge. Thus in the previous example:

*The dog is chasing the cat but it climbs the tree.*

Its parsing result indicates that ‘it’ is the subject of the verb ‘climb’. Then for ‘the dog’, ‘the cat’, and ‘the tree’, we check their associations with ‘climb’ in the knowledge base and group them in the buckets to form the SP knowledge features.

### 4.3 Baselines

Several baselines are compared in this work. The first two are conventional unsupervised ones:

- **Recent Candidate**, which simply selects the most recent noun phrase that appears in front of the target pronoun.
- **Deterministic** model (Raghunathan et al., 2010), which proposes one multi-pass sieve model with human designed rules for the coreference resolution task.

Besides the unsupervised models, we also compare with three representative supervised ones:

- **Statistical** model, proposed by Clark and Manning (2015), uses human-designed entity-level

<sup>9</sup><https://dumps.wikimedia.org/enwiki/>

<sup>10</sup>In Stanford parser results, when a verb is a linking verb (e.g., am, is), an ‘nsubj’ edge is created between its predicative and subject. Thus for this case the predicative is treated as the predicate for the subject (argument) in our study.

<sup>11</sup>If a noun phrase contains multiple words, we use the parsed result to locate its keyword and use it to represent the entire noun phrase.

Model	Third Personal Pronoun			Possessive Pronoun			All		
	P	R	F1	P	R	F1	P	R	F1
Recent Candidate	50.7	40.0	44.7	64.1	45.5	53.2	54.4	41.6	47.2
Deterministic (Raghunathan et al., 2010)	68.7	59.4	63.7	51.8	64.8	57.6	62.3	61.0	61.7
Statistical (Clark and Manning, 2015)	69.1	62.6	65.7	58.0	65.3	61.5	65.3	63.4	64.3
Deep-RL (Clark and Manning, 2016)	72.1	68.5	70.3	62.9	74.5	68.2	68.9	70.3	69.6
End2end (Lee et al., 2018)	75.1	83.7	79.2	73.9	82.1	77.8	74.8	83.2	78.8
Feature Concatenation	73.5	88.3	80.2	72.5	87.3	79.2	73.2	87.9	79.9
The Complete Model	75.4	87.9	<b>81.2</b>	74.9	87.2	<b>80.6</b>	75.2	87.7	<b>81.0</b>

Table 2: Pronoun coreference resolution performance of different models on the evaluation dataset. Precision (P), recall (R), and F1 score are reported, with the best one in each F1 column marked bold.

features between clusters and mentions for coreference resolution.

- **Deep-RL** model, proposed by Clark and Manning (2016), a reinforcement learning method to directly optimize the coreference matrix instead of the traditional loss function.
- **End2end** is the current state-of-the-art coreference model (Lee et al., 2018), which performs in an end-to-end manner and leverages both the contextual information and a pre-trained language model (Peters et al., 2018).

Note that the Deterministic, Statistical, and Deep-RL models are included in the Stanford CoreNLP toolkit<sup>12</sup>, and experiments are conducted with their provided code. For End2end, we use their released code<sup>13</sup> and replace its mention detection component with gold mentions for the fair comparison.

To clearly show the effectiveness of the proposed model, we also present a variation of our model as an extra baseline to illustrate the effect of different knowledge incorporation manner:

- **Feature Concatenation**, a simplified version of the complete model that removes the second knowledge processing layer, but directly treats all external knowledge embeddings as features and concatenates them to span representations.

#### 4.4 Implementation

Following previous work (Lee et al., 2018), we use the concatenation of the 300d GloVe embeddings (Pennington et al., 2014) and the ELMo (Peters et al., 2018) embeddings as the initial word representations. Out-of-vocabulary words are initialized with zero vectors. Hyper-parameters are

set as follows. The hidden state of the LSTM module is set to 200, and all the feed-forward networks in our model have two 150-dimension hidden layers. The default pruning threshold  $t$  for softmax pruning is set to  $10^{-7}$ . All linguistic features (plurality and AG) and external knowledge (SP) are encoded as 20-dimension embeddings.

For model training, we use cross-entropy as the loss function and Adam (Kingma and Ba, 2015) as the optimizer. All the aforementioned hyper-parameters are initialized randomly, and we apply dropout rate 0.2 to all hidden layers in the model. Our model treats a candidate as the correct reference if its predicted overall score  $F(n, p)$  is larger than 0. The model training is performed with up to 100 epochs, and the best one is selected based on its performance on the development set.

## 5 Experimental Results

Table 2 compares the performance of our model with all baselines. Overall, our model performs the best with respect to all evaluation metrics. Several findings are also observed from the results. First, manually defined knowledge and features are not enough to cover rich contextual information. Deep learning models (e.g., End2end and our proposed models), which leverage text representations for context, outperform other approaches by a great margin, especially on the recall. Second, external knowledge is highly helpful in this task, which is supported by that our model outperforms the End2end model significantly.

Moreover, the comparison between the two variants of our models is also interesting, where the final two-layer model outperforms the Feature Concatenation model. It proves that simply treating external knowledge as the feature, even though they are from the same sources, is not as effective as learning them in a joint framework. The reason

<sup>12</sup><https://stanfordnlp.github.io/CoreNLP/coref.html>

<sup>13</sup><https://github.com/kentonl/e2e-coref>

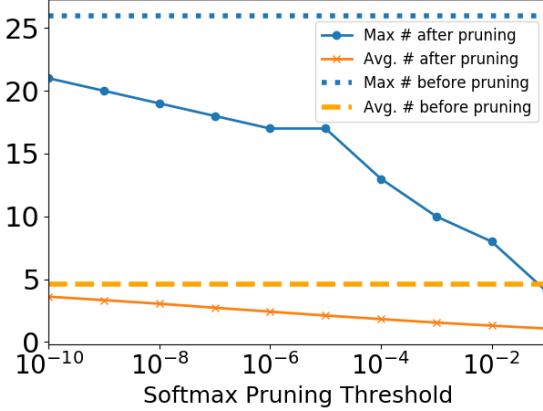


Figure 5: Effect of different thresholds on candidate numbers. Max and Average number of candidates after pruning are represented with solid lines in blue and orange, respectively. Two dashed lines indicate the max and the average number of candidates before pruning.

	F1	$\Delta F1$
The Complete Model	81.0	-
-Plurality knowledge	80.7	-0.3
-AG knowledge	80.5	-0.5
-SP knowledge	80.4	-0.6
-Knowledge Attention	80.1	-0.9

Table 3: Performance of our model with removing different knowledge sources and knowledge attention.

behind this result is mainly from the noise in the knowledge source, e.g., parsing error, incorrectly identified relations, etc. For example, the plurality of 17% noun phrases are wrongly labeled in the test data. As a comparison, our knowledge attention might contribute to alleviate such noise when incorporating all knowledge sources.

**Effect of Different Knowledge** To illustrate the importance of different knowledge sources and the knowledge attention mechanism, we ablate various components of our model and report the corresponding F1 scores on the test data. The results are shown in Table 3, which clearly show the necessity of the knowledge. Interestingly, AG contributes the most among all knowledge types, which indicates that potentially more cases in the evaluation dataset demand on the AG knowledge than others. More importantly, the results also prove the effectiveness of the knowledge attention module, which contributes to the performance gap between our model and the Feature Concatenation one.

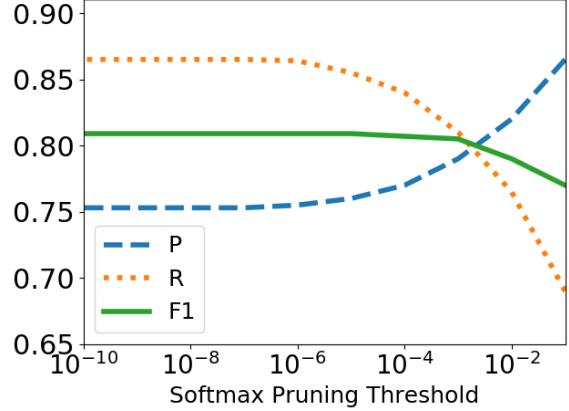


Figure 6: Effect of different pruning thresholds on model performance. With the threshold increasing, the precision increases while the recall and F1 drop.

**Effect of Different Pruning Thresholds** We try different thresholds  $t$  for the softmax pruning in selecting reliable candidates. The effects of different thresholds on reducing candidates and overall performance are shown in Figure 5 and 6 respectively. Along with the increase of  $t$ , both the max and the average number of pruned candidates drop quickly, so that the space complexity of the model can be reduced accordingly. Particularly, there are as much as 80% candidates can be filtered out when  $t = 10^{-1}$ . Meanwhile, when referring to Figure 6, it is observed that the model performs stable with the decreasing of candidate numbers. Not surprisingly, the precision rises when reducing candidate numbers, yet the recall drops dramatically, eventually results in the drop of F1. With the above observations, the reason we set  $t = 10^{-7}$  as the default threshold is straightforward: on this value, one-third candidates are pruned with almost no influence on the model performance in terms of precision, recall, and the F1 score.

## 6 Case Study

To further demonstrate the effectiveness of incorporating knowledge into pronoun coreference resolution, two examples are provided for detailed analysis. The prediction results of the End2end model and our complete model are shown in Table 4. There are different challenges in both examples. In Example A, ‘Jesus’, ‘man’, and ‘my son’ are all similar (male) noun phrases matching the target pronoun ‘He’. The End2end model predicts all of them to be correct references because their context provides limited help in dis-

	Example A	Example B
Sentences	... (A large group of people) met (Jesus). (A man in the group) shouted to him: “(Teacher), please come and look at ( <u>my son</u> ). <b>He</b> is the only child I have” ...	... (My neighbor) told me that there was ( <u>an accident</u> ), and everyone else was intact, except (his father), who was in (hospital) for fractures. I comforted him first and asked (my friend) to rush me to (the hospital). (My neighbor) showed me the police report at (the hospital), which indicated <b>it</b> was all my neighbor’s fault. ...
Pronoun Candidate NPs	<b>He</b> A large group of people, Jesus, A man in the group, Teacher, my son.	<b>it</b> My friend, an accident, his father, hospital, my friend, the hospital, My neighbor, the hospital.
End2end	Jesus, A man in the group, <u>my son</u>	None
Our Model	<u>my son</u>	<u>an accident</u>

Table 4: The comparison of End2end and our model on two examples drawn from the test data. Pronouns are marked as blue bold font. Correct references are indicated in green underline font and other candidates are indicated with brackets. ‘None’ refers to that none of the candidates is predicated as the correct reference.

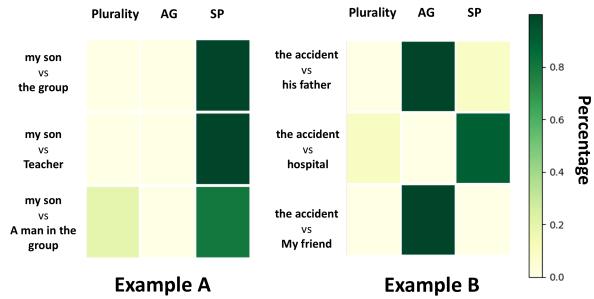


Figure 7: Heatmaps of knowledge attention for two examples, where in each example the knowledge attention weights of the correct references against other candidates are illustrated. Darker color refers to higher weight on the corresponding knowledge type.

tinguishing them. In Example B, the distance between ‘an accident’ and the pronoun ‘it’ is too far. As a result, the ‘None’ result from the End2end model indicates that the contextual information is not enough to make the decision. As a comparison, in our model, integrating external knowledge can help to solve such challenges, e.g., for Example A, SP knowledge helps when Plurality and AG cannot distinguish all candidates.

To clearly illustrate how our model leverages the external knowledge, we visualize the knowledge attention of the correct reference against other candidates<sup>14</sup> via heatmaps in Figure 7. Two interesting observations are drawn from the visualization. First, given two candidates, if they are significantly different in one feature, our model tends to pay more attention to that feature. Take AG as an example, in Example A, the AG features of all candidates consistently match the pronoun

‘he’ (all male/neutral). Thus the comparison between ‘my son’ and all candidates pay no attention to the AG feature. While in Example B, the target pronoun ‘it’ cannot describe human, thus ‘father’ and ‘friend’ are 0 on the AG feature while ‘hospital’ and ‘accident’ are 1. As a result, the attention module emphasizes AG more than other knowledge types. Second, The importance of SP is clearly shown in these examples. In example A, Plurality and AG features cannot help, the attention module weights higher on SP because ‘son’ appears 100 times as the argument of the parsed predicate ‘child’ in the SP knowledge base, while other candidates appear much less at that position. In example B, as mentioned above, once AG helps filtering ‘hospital’ and ‘accident’, SP plays an important role in distinguishing them because ‘accident’ appears 26 times in the SP knowledge base as the argument of the ‘fault’ from the results of the parser, while ‘hospital’ never appears at that position.

## 7 Related Work

Coreference resolution is a core task for natural language understanding, where it detects mention span and identifies coreference relations among them. As demonstrated in (Lee et al., 2017), mention detection and coreference prediction are the two major focuses of the task. Different from the general coreference task, pronoun coreference resolution has its unique challenge since the semantics of pronouns are often not as clear as normal noun phrases, in general, how to leverage the context and external knowledge to resolve the coreference for pronouns becomes its focus (Hobbs,

<sup>14</sup>Only candidates entered the second layer are considered.

1978; Rahman and Ng, 2011; Emami et al., 2018).

In previous work, external knowledge including manually defined rules (Hobbs, 1978; Ng, 2005), such as number/gender requirement of different pronouns, and world knowledge (Rahman and Ng, 2011), such as selectional preference (Wilks, 1975; Zhang and Song, 2018) and eventuality knowledge (Zhang et al., 2019), have been proved to be helpful for pronoun coreference resolution. Recently, with the development of deep learning, Lee et al. (2017) proposed an end-to-end model that learns contextual information with an LSTM module and proved that such knowledge is helpful for coreference resolution when the context is properly encoded. The aforementioned two types of knowledge have their own advantages: the contextual information covers diverse text expressions that are difficult to be predefined while the external knowledge is usually more precisely constructed and able to provide extra information beyond the training data. Different from previous work, we explore the possibility of joining the two types of knowledge for pronoun coreference resolution rather than use only one of them. To the best of our knowledge, this is the first attempt that uses deep learning model to incorporate contextual information and external knowledge for pronoun coreference resolution.

## 8 Conclusion

In this paper, we proposed a two-layer model for pronoun coreference resolution, where the first layer encodes contextual information and the second layer leverages external knowledge. Particularly, a knowledge attention mechanism is proposed to selectively leverage features from different knowledge sources. As an enhancement to existing methods, the proposed model combines the advantage of conventional feature-based models and deep learning models, so that context and external knowledge can be synchronously and effectively used for this task. Experimental results and case studies demonstrate the superiority of the proposed model to state-of-the-art baselines. Since the proposed model adopted an extensible structure, one possible future work is to explore the best way to enhance it with more complicated knowledge resources such as knowledge graphs.

## Acknowledgements

This paper was partially supported by the Early Career Scheme (ECS, No.26206717) from Research Grants Council in Hong Kong. In addition, Hongming Zhang has been supported by the Hong Kong Ph.D. Fellowship and the Tencent Rhino-Bird Elite Training Program. We also thank the anonymous reviewers for their valuable comments and suggestions that help improving the quality of this paper.

## References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.
- Eugene Charniak and Micha Elsner. 2009. Em works for pronoun anaphora resolution. In *Proceedings of EACL 2009*, pages 148–156.
- Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of ACL-IJCNLP 2015*, volume 1, pages 1405–1415.
- Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of EMNLP 2016*, pages 2256–2262.
- Richard J Edens, Helen L Gaylard, Gareth JF Jones, and Adenike M Lam-Adesina. 2003. An investigation of broad coverage automatic pronoun resolution for information retrieval. In *Proceedings of SIGIR 2003*, pages 381–382. ACM.
- Kate Ehrlich. 1981. Search and inference strategies in pronoun resolution: An experimental study. In *Proceedings of ACL 1981*, pages 89–93.
- Ali Emami, Paul Trichelaire, Adam Trischler, Kaeer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2018. The hard-core coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. *arXiv preprint arXiv:1811.01747*.
- Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

- Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of EMNLP 2017*, pages 188–197.
- Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of NAACL-HLT 2018*, pages 687–692.
- Dingcheng Li, Tim Miller, and William Schuler. 2011. A pronoun anaphora resolution system based on factorial hidden markov models. In *Proceedings of ACL 2011*, pages 1169–1178.
- Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of ACL 1998*, pages 869–875.
- Ruslan Mitkov. 2014. *Anaphora resolution*. Routledge.
- 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.
- Testuya Nasukawa. 1994. Robust method of pronoun resolution using full-text information. In *Proceedings of CCL 1994*, pages 1157–1163.
- Vincent Ng. 2005. Supervised ranking for pronoun resolution: Some recent improvements. In *Proceedings of EMNLP 2005*, volume 20, page 1081.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*, pages 2227–2237.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Proceedings of EMNLP 2012*, pages 1–40.
- Karthik Raghunathan, Heeyoung Lee, Sudarshan Ranganathan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of EMNLP 2010*, pages 492–501.
- Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP 2009*, pages 968–977.
- Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of ACL 2011*, pages 814–824.
- Yan Song, Chia-Jung Lee, and Fei Xia. 2017. Learning word representations with regularization from prior knowledge. In *Proceedings of CoNLL 2017*, pages 143–152.
- Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of NAACL-HLT 2018*, pages 175–180.
- Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Jevzek. 2007. Two uses of anaphora resolution in summarization. *Information Processing & Management*, 43(6):1663–1680.
- Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL 2003*, pages 168–175.
- Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.
- Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2019. Aser: A large-scale eventuality knowledge graph. *arXiv preprint arXiv:1905.00270*.
- Hongming Zhang and Yangqiu Song. 2018. A distributed solution for winograd schema challenge. In *Proceedings of ICMLC 2018*, pages 322–326.
- Yingyi Zhang, Jing Li, Yan Song, and Chengzhi Zhang. 2018. Encoding conversation context for neural keyphrase extraction from microblog posts. In *Proceedings of NAACL-HLT 2018*, pages 1676–1686.