# Knowledge-aware Pronoun Coreference Resolution

**Hongming Zhang**[♣]*, **Yan Song**[♠], **Yangqiu Song**[♣], **and Dong Yu**[♠]
[♣]Department of CSE, The Hong Kong University of Science and Technology
[♠]Tencent AI Lab
hzhangal@cse.ust.hk, clksong@gmail.com, yqsong@cse.ust.hk, dyu@tencent.com

## Abstract

Resolving pronoun coreference requires knowledge support, especially for particular domains (e.g., medicine). In this paper, we explore how to leverage different types of knowledge to better resolve pronoun coreference with a neural model. To ensure the generalization ability of our model, we directly incorporate knowledge in the format of triplets, which is the most common format of modern knowledge graphs, instead of encoding it with features or rules as that in conventional approaches. Moreover, since not all knowledge is helpful in certain contexts, to selectively use them, we propose a knowledge attention module, which learns to select and use informative knowledge based on contexts, to enhance our model. Experimental results on two datasets from different domains prove the validity and effectiveness of our model, where it outperforms state-of-the-art baselines by a large margin. Moreover, since our model learns to use external knowledge rather than only fitting the training data, it also demonstrates superior performance to baselines in the cross-domain setting.

## 1 Introduction

Being an important human language phenomenon, coreference brings simplicity for human languages while introducing a huge challenge for machines to process, especially for pronouns, which are hard to be interpreted owing to their weak semantic meanings (Ehrlich, 1981). As one challenging yet vital subtask of the general coreference resolution, pronoun coreference resolution (Hobbs, 1978) is to find the correct reference for a given pronominal anaphor in the context and has showed its importance in many natural language processing (NLP) tasks, such as machine translation (Mitkov et al., 1995), dialog systems (Strube and Müller, 2003), information extraction (Edens et al., 2003), and summarization (Steinberger et al., 2007), etc.

| | Example A | Example B |
|---|---|---|
| Sentence | The apple on the table looks great and I want to eat **it**. | Yesterday, the patient took the CT scan in the hospital and **it** showed that she had recovered. |
| Pronoun | it | it |
| Answer | The apple | the CT scan |
| Knowledge | We can eat apples but we cannot eat a table. | A 'test' shows results to patients; 'the CT scan' is a medical test. |

Table 1: Demonstration of two pronoun coreference examples, which require complex knowledge (explained in the table) to resolve. Pronouns and their corresponding mentions are marked in bold red and underline blue fonts, respectively.

In general, to resolve pronoun coreferences, one needs intensive knowledge support. As shown in Table 1, answering the first question requires the knowledge on which object can be eaten (apple v.s. table), while the second question requires the knowledge that the CT scan is a test (not the hospital) and only tests can show something. Previously, rule-based (Hobbs, 1978; Nasukawa, 1994; Mitkov, 1998; Zhang et al., 2019a) and feature-based (Ng, 2005; Charniak and Elsner, 2009; Li et al., 2011) supervised models were proposed to integrate knowledge to this task. However, while easy to incorporate external knowledge, these traditional methods faced the problem of no effective representation learning models can handle such complex knowledge. Later, end-to-end solutions with neural models (Lee et al., 2017, 2018) achieved good performance on the general coreference resolution task. Although such algo-

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



*Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 867–876
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

rithms can effectively incorporate contextual information from large-scale external unlabeled data into the model, they are insufficient to incorporate existing complex knowledge into the representation for covering all the knowledge one needs to build a successful pronoun coreference system. In addition, overfitting is always observed on deep models, whose performance is thus limited in cross-domain scenarios and restricts their usage in real applications (Liu et al., 2018, 2019). Recently, a joint model (Zhang et al., 2019b) was proposed to connect the contextual information and human-designed features together for pronoun coreference resolution task (with gold mention support) and achieved the state-of-the-art performance. However, their model still requires the complex features designed by experts, which is expensive and difficult to acquire, and requires the support of the gold mentions.

To address the limitations of the aforementioned models, in this paper, we propose a novel end-to-end model that learns to resolve pronoun coreferences with general knowledge graphs (KGs). Different from conventional approaches, our model does not require to use featurized knowledge. Instead, we directly encode knowledge triplets, the most common format of modern knowledge graphs, into our model. In doing so, the learned model can be easily applied across different knowledge types as well as domains with adopted KG. Moreover, to address the knowledge matching issue, we propose a knowledge attention module in our model, which learns to select the most related and helpful knowledge triplets according to different contexts. Experiments conducted on general (news) and in-domain (medical) cases shows that the proposed model outperforms all baseline models by a great margin. Additional experiments with the cross-domain setting further illustrate the validity and effectiveness of our model in leveraging knowledge smartly rather than fitting with limited training data[1]. To summarize, this paper makes the following contributions:

1. We explore how to resolve pronoun coreferences with KGs, which outperforms all existing models by a large margin on datasets from two different domains.
2. We propose a knowledge attention module, which helps to select the most related and helpful knowledge from different KGs.
3. We evaluate the performance of different pronoun coreference models in a cross-domain setting and show that our model has better generalization ability than state-of-the-art baselines.

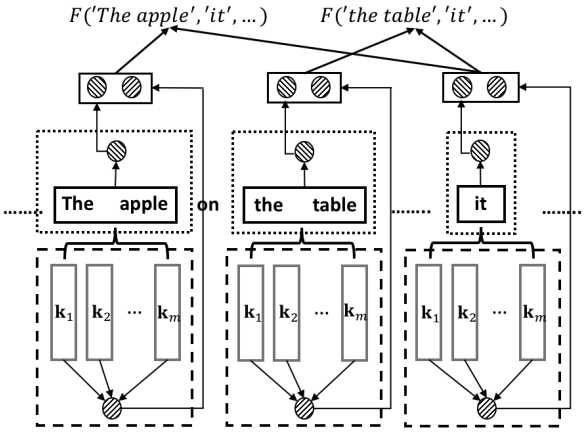


Figure 1: The overall framework of our approach to pronoun coreference resolution with KGs. $\mathbf{k}_1$,...,$\mathbf{k}_m$ represent the retrieved knowledge for each span in the black boxes. Dotted box represents the span representation module, which generates a contextual representation for each span. Dashed box represents the knowledge selection module, which selects appropriate knowledge based on the context and generates an overall knowledge representation for each span. $F(\cdot)$ is the overall coreference scoring function.

## 2 The Task

Given a text $D$, which contains a pronoun $p$, the goal is to identify all the mentions that $p$ refers to. We denote the correct mentions $p$ refers to as $c \in \mathcal{C}$, where $\mathcal{C}$ is the correct mention set. Similarly, each candidate span is denoted as $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of all candidate spans. Note that in the case where no golden mentions are annotated, all possible spans in $D$ are used to form $\mathcal{S}$. To exploit knowledge, we denote the knowledge set as $\mathcal{G}$, instantiated by multiple knowledge triplets[2]. The task is thus to identify $\mathcal{C}$ out of $\mathcal{S}$ with the support of $\mathcal{G}$. Formally, it optimizes

$$\mathcal{J} = \frac{\sum_{c \in \mathcal{C}} e^{F(c,p,\mathcal{G},D)}}{\sum_{s \in \mathcal{S}} e^{F(s,p,\mathcal{G},D)}}, \quad (1)$$

where $F(\cdot)$ is the overall scoring function[3] of $p$ referring to $s$ in $D$ with $\mathcal{G}$. The details of $F$ are illustrated in the following section.

[1] All code and data are available at: https://github.com/HKUST-KnowComp/Pronoun-Coref-KG.

[2] Each triplet contains a head, a tail, and a relation from the head to the tail.

[3] We omit $\mathcal{G}$ and $D$ in the rest of this paper for simplicity.

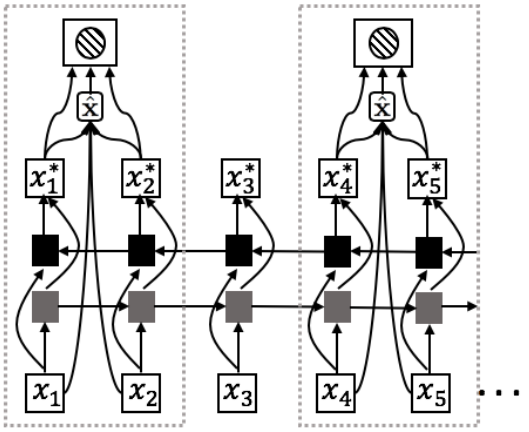

Figure 2: The structure of the span representation module. BiLSTM and attention are employed to encode the contextual information.

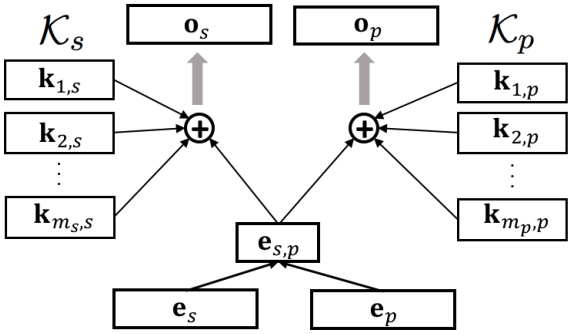

Figure 3: The structure of the knowledge attention module. The joint representation of the candidate span and pronoun is used to select knowledge for $s$ and $p$.

## 3 Model

The overall framework of our model is shown in Figure 1. There are several layers in it. At the bottom, we encode all mention spans ($s$) and pronouns ($p$) into embeddings so as to incorporate contextual information. In the middle layer, for each pair of ($s$, $p$), we use their embeddings to select the most helpful knowledge triplets from $\mathcal{G}$ and generate the knowledge representation of $s$ and $p$. At the top layer, we concatenate the textual and knowledge representation as the final representation of each $s$ and $p$, and then use this representation to predict whether there exists the coreference relation between them.

### 3.1 Span Representation

Contextual information is crucial to distinguish the semantics of a word or phrase, especially for text representation learning (Song et al., 2018; Song and Shi, 2018). In this work, a standard bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber, 1997) model is used to encode each span with attentions (Bahdanau et al., 2014), which is similar to the one used in Lee et al. (2017). The structure is shown in Figure 2. Let initial word embeddings in a span $s_i$ be denoted as $\mathbf{x}_1, ..., \mathbf{x}_T$ and their encoded representation be $\mathbf{x}^*_1, ..., \mathbf{x}^*_T$. The weighted embeddings of each span $\hat{\mathbf{x}}_i$ is obtained by

$$\hat{\mathbf{x}}_i = \sum_{t=1}^{T} a_t \cdot \mathbf{x}_t, \tag{2}$$

where $a_t$ is the inner-span attention computed by

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

where $\alpha_t$ is a standard feed-forward neural network[4] $\alpha_t = NN_\alpha(\mathbf{x}^*_t)$.

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

$$\mathbf{e}_i = [\mathbf{x}^*_{start}, \mathbf{x}^*_{end}, \hat{\mathbf{x}}_i, \phi(i)]. \tag{4}$$

Thus the span representation of $s$ and $p$ are marked as $\mathbf{e}_s$ and $\mathbf{e}_p$, respectively.

### 3.2 Knowledge Representation

For each candidate span $s$ and the target pronoun $p$, different knowledge from a KG can be extracted with various methods. For simplicity and generalization consideration, we use the string match in our model for knowledge extraction. Specifically, for each triplet $t \in \mathcal{G}$ where the head and tail of $t$ are both lists of words, if its head is the same as the string of $s$, we consider it to be a related triplet. Therefore, we encode the information of $t$ by the averaging embeddings of all words in its tail. For example, if $s$ is 'the apple' and the knowledge triplet ('the apple', *IsA*, 'healthy food') is found by searching the KG, we represent this relation from the averaged embeddings of 'healthy' and 'food'. Consequently, for $s$ and $p$, we denote their retrieved knowledge set as $\mathcal{K}_s$ and $\mathcal{K}_p$ respectively, where $\mathcal{K}_s$ contains $m_s$ related knowledge embeddings $\mathbf{k}_{1,s}$, $\mathbf{k}_{2,s}$, ..., $\mathbf{k}_{m_s,s}$ and $\mathcal{K}_p$ contains $m_p$ of them $\mathbf{k}_{1,p}$, $\mathbf{k}_{2,p}$, ..., $\mathbf{k}_{m_p,p}$.

[4] We use $NN$ to present feed-forward neural networks.

To incorporate the aforementioned knowledge embeddings into our model, we face a challenge that there are a huge number of such embeddings while most of them are useless in certain contexts. To solve it, a knowledge attention module is proposed to select the appropriate knowledge.

For each pair of ($s$, $p$), as shown in Figure 3, we first concatenate $\mathbf{e}_s$ and $\mathbf{e}_p$ to get the overall (span, pronoun) representation $\mathbf{e}_{s,p}$, which is used to select knowledge for both $s$ and $p$. Taking that for $s$ as example, we compute the weight of each $\mathbf{k}_i \in \mathcal{K}_s$ by

$$w_i = \frac{e^{\beta_{\mathbf{k}_i}}}{\sum_{\mathbf{k}_j \in \mathcal{K}_s} e^{\beta_{\mathbf{k}_j}}}, \tag{5}$$

where $\beta_{\mathbf{k}} = NN_\beta([\mathbf{e}_{s,p}, \mathbf{k}])$. As a result, the knowledge of $s$ is summed by

$$\mathbf{o}_s = \sum_{\mathbf{k}_i \in \mathcal{K}_s} w_i \cdot \mathbf{k}_i. \tag{6}$$

to represent the overall knowledge for $s$. A similar process is also conducted for $p$ with its knowledge representation $\mathbf{o}_p$.

### 3.3 Scoring

The final score of each pair ($s$, $p$) is computed by

$$F(s, p) = f_m(s) + f_c(s, p), \tag{7}$$

where $f_m(s) = NN_m([\mathbf{e}_s, \mathbf{o}_s])$ is the scoring function for $s$ to be a valid mention and $f_c(s, p) = NN_c([\mathbf{e}_n, \mathbf{o}_n, \mathbf{e}_p, \mathbf{o}_p, \mathbf{e}_n \odot \mathbf{e}_p, \mathbf{o}_n \odot \mathbf{o}_p])$ is the scoring function to identify whether there exists a coreference relation from $p$ to $s$, with $\odot$ denoting element-wise multiplication.

After getting the coreference score for all mention spans, we adopt a softmax selection on the most confident candidates for the final prediction, which is formulated as

$$\hat{F}(s, p) = \frac{e^{F(s,p)}}{\sum_{s_i \in \mathcal{S}} e^{F(s_i,p)}}. \tag{8}$$

where candidates with score $\hat{F}$ higher than a threshold $t$ are selected.

## 4 Experiments

Experiments are illustrated in this section.

| Dataset | | TP | Poss | Dem | All |
|---|---|---|---|---|---|
| CoNLL | train | 21,828 | 7,749 | 2,229 | 31,806 |
| | dev | 2,518 | 1,007 | 222 | 3,747 |
| | test | 2720 | 1,037 | 321 | 4,078 |
| i2b2 | train | 2,024 | 685 | 270 | 2,979 |
| | test | 1,244 | 367 | 166 | 1,777 |
| Overall | | 30,334 | 10,845 | 3,208 | 44,387 |

Table 2: Statistics of the two datasets. 'TP', 'Poss', and 'Dem' refer to third personal, possessive, and demonstrative pronouns, respectively.

### 4.1 Datasets

Two datasets are used in our experiments, where they are from two different domains:

- **CoNLL:** The CoNLL-2012 shared task (Pradhan et al., 2012) corpus, which is a widely used dataset selected from the Ontonotes 5.0[5].
- **i2b2:** The i2b2 shared task dataset (Uzuner et al., 2012), consisting of electronic medical records from two different organizations, namely, Partners HealthCare (Part) and Beth Israel Deaconess medical center (Beth). All records have been fully de-identified and manually annotated with coreferences.

We split the datasets into different proportions based on their original settings. Three types of pronouns are considered in this paper following Ng (2005), i.e., third personal pronoun (e.g., *she*, *her*, *he*, *him*, *them*, *they*, *it*), possessive pronoun (e.g., *his*, *hers*, *its*, *their*, *theirs*), and demonstrative pronoun (e.g., *this*, *that*, *these*, *those*). Table 2 reports the number of the three types of pronouns and the overall statistics of the experiment datasets with proportion splittings. Following conventional approaches (Ng, 2005; Li et al., 2011), for each pronoun, we consider its candidate mentions from the previous two sentences and the current sentence it belongs to. According to our selection range of the candidate mentions, each pronoun in the CoNLL data and i2b2 data has averagely 1.3 and 1.4 correct references, respectively.

### 4.2 Knowledge Resources

As mentioned in previous sections, our model is designed to leverage general KGs, where it takes triplets as the input of knowledge representations. For all knowledge resources, we format them as triplets and merge them together to obtain the final

[5]https://catalog.ldc.upenn.edu/LDC2013T19

knowledge set. Different knowledge resources are introduced as follows.

**Commonsense knowledge graph (OMCS).** We use the largest commonsense knowledge base, the open mind common sense (OMCS) (Singh, 2002) in this paper. OMCS contains 600K crowd-sourced commonsense triplets such as *(food, UsedFor, eat)* and *(wind, CapableOf, blow to east)*. All relations in OMCS are human-defined and we select those highly-confident ones (confidence score larger than 2) to form the OMCS KG, with 62,730 triplets.

**Medical concepts (Medical-KG).** Being part of the i2b2 contest, the related knowledge about medical concepts such as *(the CT scan, is, test)* and *(intravenous fluids, is, treatment)* are provided. The annotated triplets are used as the medical concept KG, which contains 22,234 triplets.

**Linguist features (Ling).** In addition to manually annotated KGs, we also consider linguist features, i.e., plurality and animacy & gender (AG), as one important knowledge resources. Stanford parser[6] is employed to generate plurality, animacy, and gender markups for all the noun phrases, so as to automatically generate linguistic knowledge (in the form of triplets) for our data. Specifically, the plurality feature denotes each $s$ and $p$ to be singular or plural. The animacy & gender (AG) feature denotes whether the $n$ or $p$ is a living object, and being male, female, or neutral if it is alive. For example, a mention 'the girls' is labeled as *plural* and *female*; we use triplets *('the girls', plurality, Plural)* and *('the girls', AG, female)* to represent them. As a result, we have 40,149 and 40,462 triplets for plurality and AG, respectively.

**Selectional Preference (SP).** Selectional preference (Hobbs, 1978) knowledge is employed as the last knowledge resource, which is the semantic constraint for word usage. SP generally refers to that, given a predicate (e.g., verb), people have the preference for the argument (e.g., its object or subject) connected. To collect SP knowledge, we first parse the English Wikipedia[7] with the Stanford parser and extract all dependency edges in the format of *(predicate, argument, relation, number)*, where predicate is the governor and argument the dependent in each dependency edge[8]. Following (Resnik, 1997), each potential SP pair is measured by a posterior probability

$$P_r(a|p) = \frac{Count_r(p,a)}{Count_r(p)}, \tag{9}$$

where $Count_r(p)$ and $Count_r(p,a)$ refer to how many times $p$ and the predicate-argument pair ($p$, $a$) appear in the relation $r$, respectively. In our experiment, if $P_r(a|p) > 0.1$ and $Count_r(p,a) > 10$, we consider the triplet ($p$, $r$, $a$) (e.g., *('dog', nsubj, 'barks')*) a valid SP relation. Finally, we select two SP relations, *nsubj* and *dobj*, to form the SP knowledge graph, including 17,074 and 4,536 frequent predicate-argument pairs for *nsubj* and *dobj*, respectively.

### 4.3 Baselines

Several baselines are compared in this paper, including three widely used pre-trained models:

- **Deterministic** model (Raghunathan et al., 2010), which is an unsupervised model and leverages manual rules to detect coreferences.
- **Statistical** model (Clark and Manning, 2015), which is a supervised model and trained on manually crafted entity-level features between clusters and mentions.
- **Deep-RL** model (Clark and Manning, 2016), which uses reinforcement learning to directly optimize the coreference matrix instead of the loss function of supervised learning.

The above models are included in the Stanford CoreNLP toolkit[9]. We also include a state-of-the-art end-to-end neural model as one of our baselines:

- **End2end** (Lee et al., 2018), which is the current state-of-the-art model performing in an end-to-end manner and leverages both contextual information and a pre-trained language model (Peters et al., 2018).

We use their released code[10]. In addition, to show the importance of incorporating knowledge, we also experiment with two variations of our model:

- **Without KG** removes the KG component and keeps all other components in the same setting as that in our complete model.

[6] https://stanfordnlp.github.io/CoreNLP/

[7] https://dumps.wikimedia.org/enwiki/

[8] In the Stanford parser, an 'nsubj' edge is created between its predictive and subject when a verb is a linking verb (e.g., am, is); the predicative is thus treated as the predicate for the subject (argument) in this paper.

[9] https://stanfordnlp.github.io/CoreNLP/coref.html

[10] https://github.com/kentonl/e2e-coref

| Model | Third Personal | | | Possessive | | | Demonstrative | | | All | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Deterministic | 25.5 | 58.9 | 35.6 | 22.9 | 64.3 | 33.8 | 3.4 | 5.7 | 4.2 | 23.4 | 57.0 | 33.4 |
| Statistical | 25.8 | 62.1 | 36.5 | 28.9 | 64.9 | 40.0 | 9.8 | 6.3 | 7.6 | 25.4 | 59.3 | 36.5 |
| Deep-RL | 78.6 | 63.9 | 70.5 | 73.3 | 68.9 | 71.0 | 3.7 | 2.9 | 5.5 | 76.4 | 61.2 | 68.0 |
| End2end | 70.6 | 75.7 | 73.1 | 73.0 | 76.2 | 74.6 | 58.4 | 17.6 | 27.0 | 71.1 | 72.1 | 71.6 |
| Without KG | 78.2 | 72.4 | 75.2 | 80.0 | 66.4 | 72.6 | 46.7 | 62.5 | 53.4 | 75.7 | 70.1 | 72.8 |
| Without Attention | 76.6 | 77.9 | 77.2 | 79.0 | 73.5 | 76.2 | 42.4 | 72.6 | 53.5 | 73.6 | 76.4 | 74.9 |
| Our Complete Model | 78.8 | 77.8 | **78.1** | 80.7 | 72.5 | **76.4** | 45.3 | 66.7 | **53.9** | 75.9 | 75.6 | **75.7** |

(a) CoNLL

| Model | Third Personal | | | Possessive | | | Demonstrative | | | All | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Deterministic | 25.7 | 57.4 | 35.5 | 25.2 | 61.6 | 35.7 | 6.6 | 4.0 | 5.0 | 25.1 | 54.0 | 34.3 |
| Statistical | 19.3 | 35.9 | 25.1 | 25.7 | 50.5 | 34.0 | 6.7 | 4.5 | 5.4 | 20.5 | 36.6 | 26.3 |
| Deep-RL | 78.2 | 48.0 | 59.5 | 78.6 | 57.7 | 66.5 | 9.1 | 5.1 | 9.6 | 77.8 | 46.3 | 58.1 |
| End2end | 95.0 | 93.4 | 94.2 | 95.3 | 96.0 | 95.7 | 74.8 | 52.5 | 61.7 | 93.9 | 90.7 | 92.3 |
| Without KG | 96.8 | 95.9 | 96.3 | 97.1 | 97.5 | 97.3 | 66.5 | 68.2 | 67.3 | 94.3 | 94.0 | 94.2 |
| Without Attention | 96.1 | 97.2 | 96.6 | 96.3 | 98.2 | 97.2 | 66.7 | 77.8 | 71.8 | 93.4 | 95.9 | 94.6 |
| Our Complete Model | 97.5 | 96.3 | **96.9** | 98.5 | 97.8 | **98.2** | 71.9 | 72.2 | **72.0** | 95.6 | 94.7 | **95.2** |

(b) i2b2

Table 3: The performance of pronoun coreference resolution with different models on two evaluation datasets. Precision (P), recall (R), and the F1 score are reported, with the best one in each F1 column marked as bold.

- **Without Attention** removes the knowledge attention module and concatenates all the knowledge embeddings. All other components are identical as our complete model.

### 4.4 Implementation

Following the 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 for computing span representations. For knowledge triplets, we use the GloVe embeddings to encode tail words in them. Out-of-vocabulary words are initialized with zero vectors. The hidden state of the LSTM module is set to 200, and all the feed-forward networks have two 150-dimension hidden layers. The selection thresholds are set to $10^{-2}$ and $10^{-8}$ for the CoNLL and i2b2 dataset, respectively.

For model training, we use cross-entropy as the loss function and Adam (Kingma and Ba, 2014) as the optimizer. All the aforementioned hyperparameters are initialized randomly, and we apply dropout rate 0.2 to all hidden layers in the model. For the CoNLL dataset, the model training is performed with up to 100 epochs, and the best one is selected based on its performance on the development set. For the i2b2 dataset, because no dev set is provided, we train the model up to 100 epochs and use the final converged one.

### 4.5 Results

Table 3 reports the performance of all models, with the results for CoNLL and i2b2 in (a) and (b), respectively. Overall, our model outperforms all baselines on two datasets with respect to all pronoun types. There are several interesting observations. In general, the i2b2 dataset seems simpler than the CoNLL dataset, which might because that i2b2 only involves clinical narratives and its training data is highly similar to the test data. As a result, all neural models perform dramatically good, especially on the third personal and possessive pronouns. In addition, we also notice that it is more challenging for all models to resolve demonstrative pronouns (e.g., *this*, *that*) on both datasets, because such pronouns may refer to complex things and occur with low frequency.

Moreover, there are significant gaps in the performance of different models, with the following observations. First, models with manually defined rules or features, which cannot cover rich contextual information, perform poorly. In contrast, 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. Sec-

| | CoNLL | | i2b2 | |
|---|---|---|---|---|
| | F1 | $\Delta F1$ | F1 | $\Delta F1$ |
| The Complete Model | 75.7 | - | 95.2 | - |
| –OMCS | 74.8 | -0.9 | 95.1 | -0.1 |
| –Medical-KG | 74.5 | -1.2 | 94.6 | -0.6 |
| –Ling | 73.8 | -1.9 | 94.9 | -0.3 |
| –SP | 74.0 | -1.7 | 94.7 | -0.5 |

Table 4: The performance of our model with removing different knowledge resources. The F1 of each case and the difference of F1 between each case and the complete model are reported.

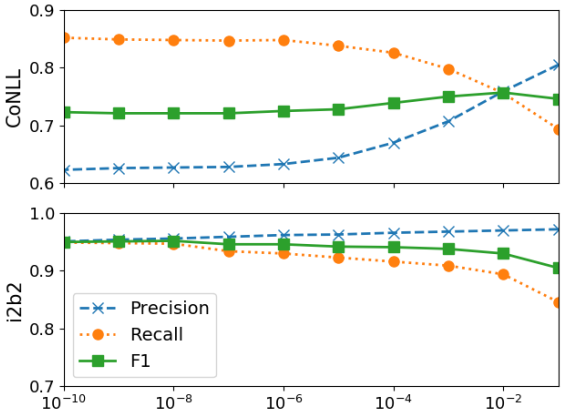


Figure 4: Effect of different softmax selection thresholds with respect to our model performance on two datasets. In general, with the threshold becoming larger, less candidates are selected, the precision thus increases while the recall drops.

ond, adding knowledge in an appropriate manner within neural models is helpful, which is supported by that our model outperforms the End2end model and the Without KG one on both datasets, especially CoNLL, where the external knowledge plays a more important role. Third, the knowledge attention module ensures our model to predict more precisely, which also results in the overall improvement on F1. To summarize, the results suggest that external knowledge is important for effectively resolving pronoun coreference, where rich contextual information determines the appropriate knowledge with a well-designed module.

## 5 Analysis

Further analysis is conducted in this section regarding the effect of different knowledge resources, model components, and settings. Details are illustrated as follows.

### 5.1 Ablation Study

We ablate different knowledge for their contributions in our model, with the results reported in Table 4. It is observed that all knowledge resources contribute to the final success of our model, where different knowledge types play their unique roles in different datasets. For example, the Ling knowledge contributes the most to the CoNLL dataset while the medical knowledge is the most important one for the medical data.

### 5.2 Effect of the Selection Threshold

We experiment with different thresholds $t$ for the softmax selection. The effects of $t$ against overall performance are shown in Figure 4. In general, with the increase of $t$, fewer candidates are selected. Therefore, the overall precision increases and the recall drops. Consider that both the precision and recall are important for resolving pronoun coreference, we select different thresholds for different datasets to ensure the balance between precision and recall. In detail, for the CoNLL dataset, we set $r = 10^{-2}$ to select the most confident predictions; and for the i2b2 dataset, we set $r = 10^{-8}$ so as to keep more predictions.

| Model | Setting | CoNLL | i2b2 |
|---|---|---|---|
| End2end | Original | 71.6 | 92.3 |
| | + Gold mention | 77.8 | 94.4 |
| Our Model | Original | 75.7 | 95.2 |
| | + Gold mention | **80.7** | **96.0** |
| (Zhang et al., 2019b) | + Gold mention | 79.9 | - |

Table 5: Influence of gold mentions. F1 scores on different test sets are reported. Adding human-annotated gold mentions help both the End2end and our model. Best performed model are indicated with the bold font.

### 5.3 Effect of Gold Mentions

The effect of adding gold mentions is shown in Table 5. Providing gold mentions to the End2end model can significantly boost its performance by 6.2 F1 and 2.1 F1 on the CoNLL and i2b2 dataset, respectively. Yet, the performance gain from gold mentions is less for our model. Such results clearly illustrate that our model is able to benefit the mention detection with the help of KG incorporation. Besides that, with the help of gold mentions, our model achieves the comparable (slightly better) performance with the context-and-knowledge model (Zhang et al., 2019b). As their features are originally designed for CoNLL, we only report the performance on CoNLL in Ta-

| Model | Training data | Test data CoNLL | i2b2 |
|---|---|---|---|
| End2end | CoNLL | 71.6 | 75.2 |
| | i2b2 | 20.0 | 92.3 |
| Our Model | CoNLL | 75.7 | 80.9 |
| | i2b2 | 42.7 | 95.2 |

Table 6: Cross-domain performance of different models. F1 on the target domain test sets are reported.

ble 5. As we also included one new challenging pronoun type, the demonstrative pronoun, the overall performance of their model is lower than the one reported in the original paper. The reason of our model being better is that more knowledge resources (e.g., OMCS) can be incorporated into our model due to its generalizable design. Moreover, it is more difficult for their method (Zhang et al., 2019b) to incorporate mention detection into the model, because in this case we need to enumerate all mention spans and generate corresponding features for all spans. which is expensive and difficult to acquire.

### 5.4 Cross-domain Evaluation

Considering that neural models are intensive data-driven and normally restricted by data nature, they are not easily applied in a cross-domain setting. However, if a model is required to perform in real applications, it has to show promising performance on those cases out of the training data. Herein we investigate the performance of different models with training and testing in different domains, with the results reported in Table 6. Overall, all models perform significantly worse if they are used cross domains. Specifically, if we train the End2end model on the CoNLL dataset and test it on the i2b2 dataset, it only achieves 75.2 F1. As a comparison, our model can achieve 80.9 F1 in the same case. This observation confirms the capability of knowledge where our model is able to handle. A similar observation is also drawn for the reversed case. However, even though our model outperforms the End2end model by 22.7 F1 from i2b2 to CoNLL, its overall performance is still poor, which might be explained by that the i2b2 is an in-domain dataset and the knowledge contained in its training data is rarely useful for the general (news) domain dataset. Nevertheless, this experiment clearly shows that the generalization ability of deep models is still crucial for building a successful coreference model, and learns to use knowledge is a promising solution to it.

| | Example A | Example B |
|---|---|---|
| Sentence | He walks into the room with one magazine and drops **it** on the couch. | ... A small area of erythema around his arm ... **This** will be treated empirically. |
| Prediction | magazine | erythema |
| Knowledge | ('magazine', *dobj*, 'drop') | ('erythema', *IsA*, 'disease') |

Table 7: The case study on two examples from the test data, i.e., A: from the CoNLL and B: from the i2b2. Pronouns and correct mentions are marked by red bold and blue underline font respectively. Knowledge triplets used for them are listed in the bottom row.

## 6 Case Study

To better illustrate the effectiveness of incorporating different knowledge in this task, two examples are provided for the case study in Table 7. In example A, our model correctly predicts that 'it' refers to the '*magazine*' rather than the '*room*', because we successfully retrieve the knowledge that compared with the 'room', the '*magazine*' is more likely to be the object of *drop*. In example B, even though the distance between 'erythema' and 'This' is relatively far[11], our model is able to determine the coreference relation between them because it successfully finds out that 'erythema' is a kind of disease, while a lot of diseases appear as the context of 'be treated' in the training data.

## 7 Related Work

Detecting mention spans in linguistic expressions and identifying coreference relations among them is a core task, namely, coreference resolution, for natural language understanding. Mention detection and coreference prediction are the two major focuses of the task as listed in Lee et al. (2017). Compared to general coreference problem, pronoun coreference resolution has its unique challenge since pronouns themselves have weak semantics meanings, which make it the most challenging sub-task in general coreference resolution. To address the unique difficulty brought by pronouns, we thus focus on resolving pronoun coreferences in this paper.

Resolving pronoun coreference relations often requires the support of manually crafted knowledge (Rahman and Ng, 2011; Emami et al., 2018),

[11] We omit the intermediate part of the long sentence in the table for a clear presentation.

especially for particular domains such as medicine (Uzuner et al., 2012) and biology (Cohen et al., 2017). Previous studies on pronoun coreference resolution incorporated external knowledge including human defined rules (Hobbs, 1978; Ng, 2005), e.g., number/gender requirement of different pronouns, domain-specific knowledge such as medical (Jindal and Roth, 2013) or biological (Trieu et al., 2018) ones, and world knowledge (Rahman and Ng, 2011), such as selectional preference (Wilks, 1975). Later, end-to-end solutions (Lee et al., 2017, 2018) were proposed to learn contextual information and solve coreferences synchronously with neural networks, e.g., LSTM. Their results proved that such knowledge is helpful when appropriately used for coreference resolution. However, external knowledge is often omitted in their models. Consider that context and external knowledge have their own advantages: the contextual information covering diverse text expressions that are difficult to be predefined while the external knowledge being usually more precisely constructed and able to provide extra information beyond the training data, one could benefit from both sides for this task. Different from previous studies, we provide a generic solution to resolving pronoun coreference with the support of knowledge graphs based on contextual modeling, where deep learning models are adopted in our work to incorporate knowledge into pronoun coreference resolution and achieve remarkably good results.

## 8 Conclusion

In this paper, we explore how to build a knowledge-aware pronoun coreference resolution model, which is able to leverage different external knowledge for this task. The proposed model is an attempt of the general solution of incorporating knowledge (in the form of KG) into the deep learning based pronoun coreference model, rather than using knowledge as features or rules in a dedicated manner. As a result, any knowledge resource presented in the format of triplets, the most widely used entry format for KG, can be consumed in our model with a proposed attention module. Experimental results on two different corpora from two domains demonstrate the superiority of the proposed model to all baselines. Moreover, as our model learns to use knowledge rather than just fitting the training data, our model achieves much better and more robust performance than state-of-the-art models in the cross-domain scenario.

## Acknowledgements

This paper was partially supported by the Early Career Scheme (ECS, No.26206717) from Research Grants Council in Hong Kong and Tencent AI Lab Rhino-Bird Focused Research Program. In addition, Hongming Zhang has been supported by the Hong Kong Ph.D. Fellowship. 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. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Eugene Charniak and Micha Elsner. 2009. Em works for pronoun anaphora resolution. In *EACL, 2009*, pages 148–156.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *ACL-IJCNLP, 2015*, volume 1, pages 1405–1415.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2256–2262.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC bioinformatics*, 18(1):372.

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 *SIGIR*, pages 381–382. ACM.

Kate Ehrlich. 1981. Search and inference strategies in pronoun resolution: An experimental study. In *ACL, 1981*, pages 89–93.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer 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.

Prateek Jindal and Dan Roth. 2013. End-to-end coreference resolution for clinical narratives. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *EMNLP, 9-11, 2017*, pages 188–197.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *NAACL-HLT*, 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.

Miaofeng Liu, Jialong Han, Haisong Zhang, and Yan Song. 2018. Domain Adaptation for Disease Phrase Matching with Adversarial Networks. In *Proceedings of the BioNLP 2018 workshop 2018*, pages 137–141.

Miaofeng Liu, Yan Song, Hongbin Zou, and Tong Zhang. 2019. Reinforced Training Data Selection for Domain Adaptation. In *Proceedings of ACL 2019*.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *ACL, 1998*, pages 869–875.

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 *CCL, 1994*, pages 1157–1163.

Vincent Ng. 2005. Supervised ranking for pronoun resolution: Some recent improvements. In *EMNLP, 2005*, volume 20, page 1081.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *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. *arXiv preprint arXiv:1802.05365*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *EMNLP, 2012*, pages 1–40.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *EMNLP, 2010*, pages 492–501.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *ACL, 2011*, pages 814–824.

Philip Resnik. 1997. Selectional preference and sense disambiguation. *Tagging Text with Lexical Semantics: Why, What, and How?*

Push Singh. 2002. The open mind common sense project. *KurzweilAI. net*.

Yan Song and Shuming Shi. 2018. Complementary Learning of Word Embeddings. In *Proceedings of IJCAI 2018*, pages 4368–4374.

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 Jeˇzek. 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 *ACL, 2003*, pages 168–175.

Long Trieu, Nhung Nguyen, Makoto Miwa, and Sophia Ananiadou. 2018. Investigating domain-specific information for neural coreference resolution on biomedical texts. In *Proceedings of the BioNLP 2018 workshop*, pages 183–188.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

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. 2019a. Aser: A large-scale eventuality knowledge graph. *arXiv preprint arXiv:1905.00270*.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019b. Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of NAACL-HLT 2019*.