Neural Coreference Resolution with Deep Biaffine Attention by Joint Mention Detection and Mention Clustering

Rui Zhang[1], Cicero Nogueira dos Santos[2], Michihiro Yasunaga[1], Bing Xiang[2], Dragomir Radev[1]

[1]Yale University, [2]IBM T. J. Watson Research Center

LILY Lab

Introduction

Coreference resolution aims to identify in a text all mentions that refer to the same real world entity. The state-of-the-art end-to-end neural coreference model considers all text spans in a document as potential mentions and learns to link an antecedent for each possible mention. In this paper, we propose to improve the end-to-end coreference resolution system by

(1) using a biaffine attention model to get antecedent scores for each possible mention,
(2) jointly optimizing the mention detection accuracy and the mention clustering log-likelihood given the mention cluster labels.

Our model achieves the state-of-the-art performance on the CoNLL-2012 Shared Task English test set.

Our method

Span Representation. We adopt the same span representation approach as in Lee et al. (2017), using bidirectional LSTMs. Then, the head-finding attention computes a score distribution over different words in a span. To construct span representation, we encode both contextual information and internal structure of spans, and a feature vector for the span size.

$$\mathbf{s}_i = [\mathbf{h}_{\text{START}(i)}, \mathbf{h}_{\text{END}(i)}, \mathbf{w}_i^{\text{head-att}}, \phi(i)]$$

Mention Scoring. The span representation is input to a feed forward network which measures if it is an entity mention using a score m(i):

$$m(i) = \mathbf{v}_{\text{m}}^{\intercal}\text{FFNN}_{\text{m}}(\mathbf{s}_i)$$

Biaffine Attention Antecedent Scoring. For antecedent scoring, we propose a biaffine attention model (Dozat and Manning, 2017) to produce distributions of possible antecedents:

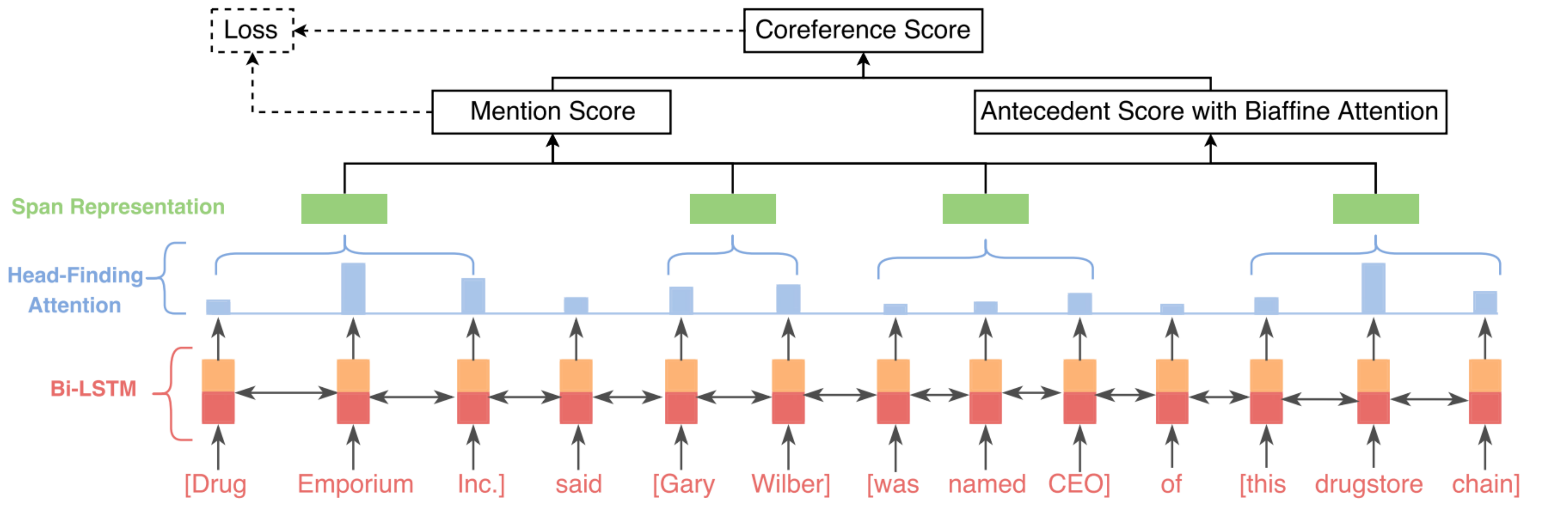

Figure 1: Model architecture. We consider all text spans up to 10-word length as possible mentions. For brevity, we only show three candidate antecedent spans (“Drug Emporium Inc.”, “Gary Wilber”, “was named CEO”) for the current span “this drugstore chain”.

$$\hat{\mathbf{s}}_{\text{i}} = \text{FFNN}_{\text{anaphora}}(\mathbf{s}_{\text{i}})$$
$$\hat{\mathbf{s}}_{\text{j}} = \text{FFNN}_{\text{antecedent}}(\mathbf{s}_{\text{j}}), 1 \leq j \leq i-1$$
$$c(i,j) = \hat{\mathbf{s}}_{\text{j}}^{\intercal}\mathbf{U}_{\text{bi}}\hat{\mathbf{s}}_{\text{i}} + \mathbf{v}_{\text{bi}}^{\intercal}\hat{\mathbf{s}}_{\text{i}}$$

Inference. The final coreference score s(i, j) forspan si and span sj consists of three terms: (1)if si is a mention, (2) if sj is a mention, (3) if sj is an antecedent for si. Furthermore, for dummy antecedent, we fix the final score to be 0:

$$s(i,j) = \begin{cases} m(i) + m(j) + c(i,j), & j \neq \epsilon \\ 0, & j = \epsilon \end{cases}$$

Joint Mention Detection and Mention Cluster. Our training data only provides gold mention cluster labels. To make best use of this information, we propose to jointly optimize the mention scoring and antecedent scoring in our loss function.

$$\mathcal{L}_{\text{cluster}}(i) = \log \frac{\sum_{j' \in \text{GOLD}(i)} \exp(s(i,j'))}{\sum_{j=\epsilon,0,\ldots,i-1} \exp(s(i,j))}$$

$$\mathcal{L}_{\text{detect}}(i) = y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)$$

$$\mathcal{L}_{\text{loss}} = -\lambda_{\text{detect}} \sum_{i=1}^{N} \mathcal{L}_{\text{detect}}(i) - \sum_{i'=1}^{N'} \mathcal{L}_{\text{cluster}}(i')$$

Experimental Results

	MUC			B^3			CEAF$_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Our work (5-model ensemble)	82.1	73.6	77.6	73.1	62.0	67.1	67.5	59.0	62.9	**69.2**
Lee et al. (2017) (5-model ensemble)	81.2	73.6	77.2	72.3	61.7	66.6	65.2	60.2	62.6	68.8
Our work (single model)	79.4	73.8	76.5	69.0	62.3	65.5	64.9	58.3	61.4	**67.8**
Lee et al. (2017) (single model)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

Table 1: Experimental results on the CoNLL-2012 Englisth test set. The F1 improvements are statistical significant with $p < 0.05$ under the paired bootstrap resample test (Koehn, 2004) compared with Lee et al. (2017).

In Table 1, we compare our model with previous state-of-the-art systems on the CoNLL-2012 Shared Task English data. We obtain the best results in all F1 metrics. Our single model achieves 67.8% F1 and our 5-model ensemble achieves 69.2% F1. In particular, compared with Lee et al. (2017), our improvement mainly results from the precision scores. This indicates that the mention detection loss does produce better mention scores and the biaffine attention more effectively determines if two spans are coreferent.

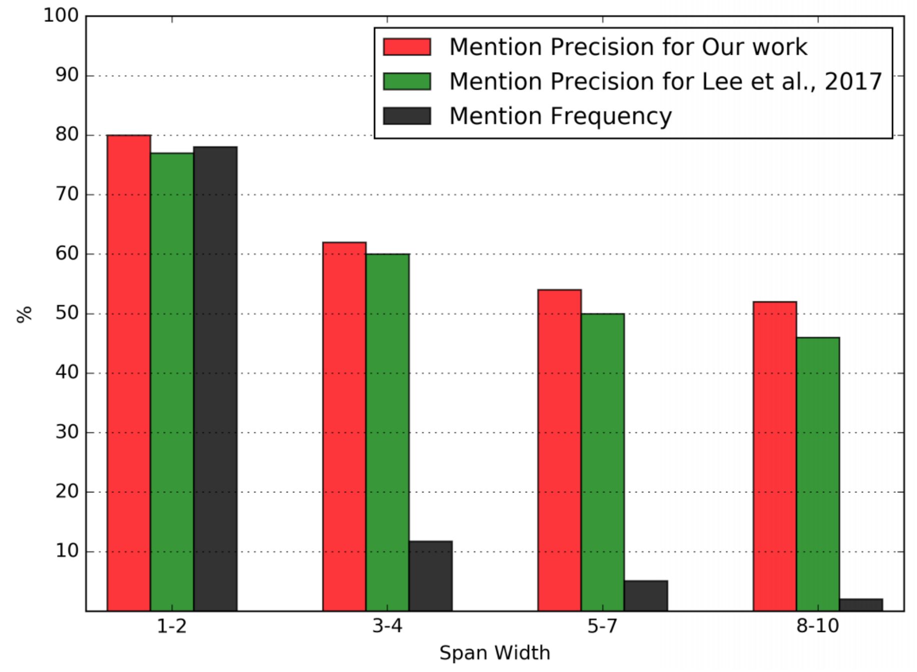

Figure 2: Mention detection subtask on development set. We plot accuracy and frequency breakdown by span widths.

Mention Detection Analysis

To further understand our model, we perform a mention detection subtask where spans with mention scores higher than 0 are considered as mentions. We show the mention detection accuracy breakdown by span widths in Figure 2. Our model indeed performs better thanks to the mention detection loss. The advantage is even clearer for longer spans which consist of 5 or more words. Here are some examples of unseen mentions in test set which our model can correctly detect but Lee et al. (2017) cannot:

(1) a suicide murder (2) Hong Kong Island
(3) a US Airforce jet carrying robotic undersea vehicles
(4) the investigation into who was behind the apparent suicide attack

This shows that our mention loss helps detection by generalizing to new mentions rather than memorizing the existing mentions in training data.