Gender Balanced Coreference Resolution

Nicholas Tan, Hongshen Zhao

Motivation

Co-reference resolution is the task of identifying all words and phrases that refer to the same entity in a corpus of text. In particular, to accomplish this task, one must find all pronouns or referring expressions and connect them to their antecedents. In natural language, these connections may be ambiguous and resolving them may rely heavily on contextual clues or implied past experiences. This makes the task especially difficult in machine understanding of natural language, where human-level performance has not yet been achieved.

Task

For example, given the following source text and ambiguous pronoun (**bolded**):
Kathleen Nott was born in Camberwell, London. Her father, Philip, was a litho- graphic printer, and her mother, Ellen, ran a boarding house in Brixton; Kathleen was their third daughter. **She** was educated at Mary Datchelor Girls' School (now closed), London, before attending King's College, London.
Our objective is to find the antecedent for **She**, which is **Kathleen**. Note that it may be ambiguous in certain interpretations of the text whether **Ellen** may have been the true antecedent.

Dataset

Our baseline evaluation used the English coreference resolution annotations from the CoNLL-2012 shared task benchmark dataset. This dataset contains 2802 training documentations, 343 development documentations, and 348 test documentations. The training documentations contain on average 454 words and a maximum of 4009 words. For our targeted study, we used the Google GAP Conference dataset, which has 8,908 coreference- labeled pairs of (ambiguous pronoun, antecedent name). The dataset has been specifically designed to represent the challenges of resolving ambiguous pronouns in a gender-balanced context. The dataset samples each excerpt from various Wikipedia articles and annotations are created with human labeling. The dataset is provided by the Google AI Language group (https://github.com/google- research-datasets/gap-coreference).

Method

We leverage work by Kenton Lee et. al. in coreference resolution using end-to-end models [2], [1]. They addressed the co-reference resolution task as the set of decisions to assign an antecedent yi for every possible span i in the document, where yi ∈ Y(i) = {ε, 1, ..., i − 1}. Several fully connected layers were appended to the predicted span outputs with deep residual skip-connections to compare the proposed span cluster with the original output from Kenton Lee et. al. These arbitrary length output sequences were padded up to a constant size before entering the fully connected layers; and after each fully connected layer, dropout was performed. Each of the layers used a ReLU activation, except for the final layer, which used a sigmoid to output a probability between 0 and 1. The layer sizes were chosen to continually decrease until one node was reached, which would be used to predict the TRUE or FALSE labeling accompanying each example in the Google GAP dataset. The loss function was thus edited to be the binary cross entropy loss instead of the previous softmax loss.

Span Representations (LSTM): To compute vector representations of each span, bidirectional LSTMs are used to encode every word in the context.

$$\mathbf{f}_{t,\delta} = \sigma(\mathbf{W}_f[\mathbf{x}_t, \mathbf{h}_{t+\delta,\delta}] + \mathbf{b}_i)$$
$$\mathbf{o}_{t,\delta} = \sigma(\mathbf{W}_o[\mathbf{x}_t, \mathbf{h}_{t+\delta,\delta}] + \mathbf{b}_o)$$
$$\widetilde{\mathbf{c}}_{t,\delta} = \tanh(\mathbf{W}_c[\mathbf{x}_t, \mathbf{h}_{t+\delta,\delta}] + \mathbf{b}_c)$$
$$\mathbf{c}_{t,\delta} = \mathbf{f}_{t,\delta} \circ \widetilde{\mathbf{c}}_{t,\delta} + (1 - \mathbf{f}_{t,\delta}) \circ \mathbf{c}_{t+\delta,\delta}$$
$$\mathbf{h}_{t,\delta} = \mathbf{o}_{t,\delta} \circ \tanh(\mathbf{c}_{t,\delta})$$
$$\mathbf{x}^* = [\mathbf{h}_{t,1}, \mathbf{h}_{t,-1}]$$

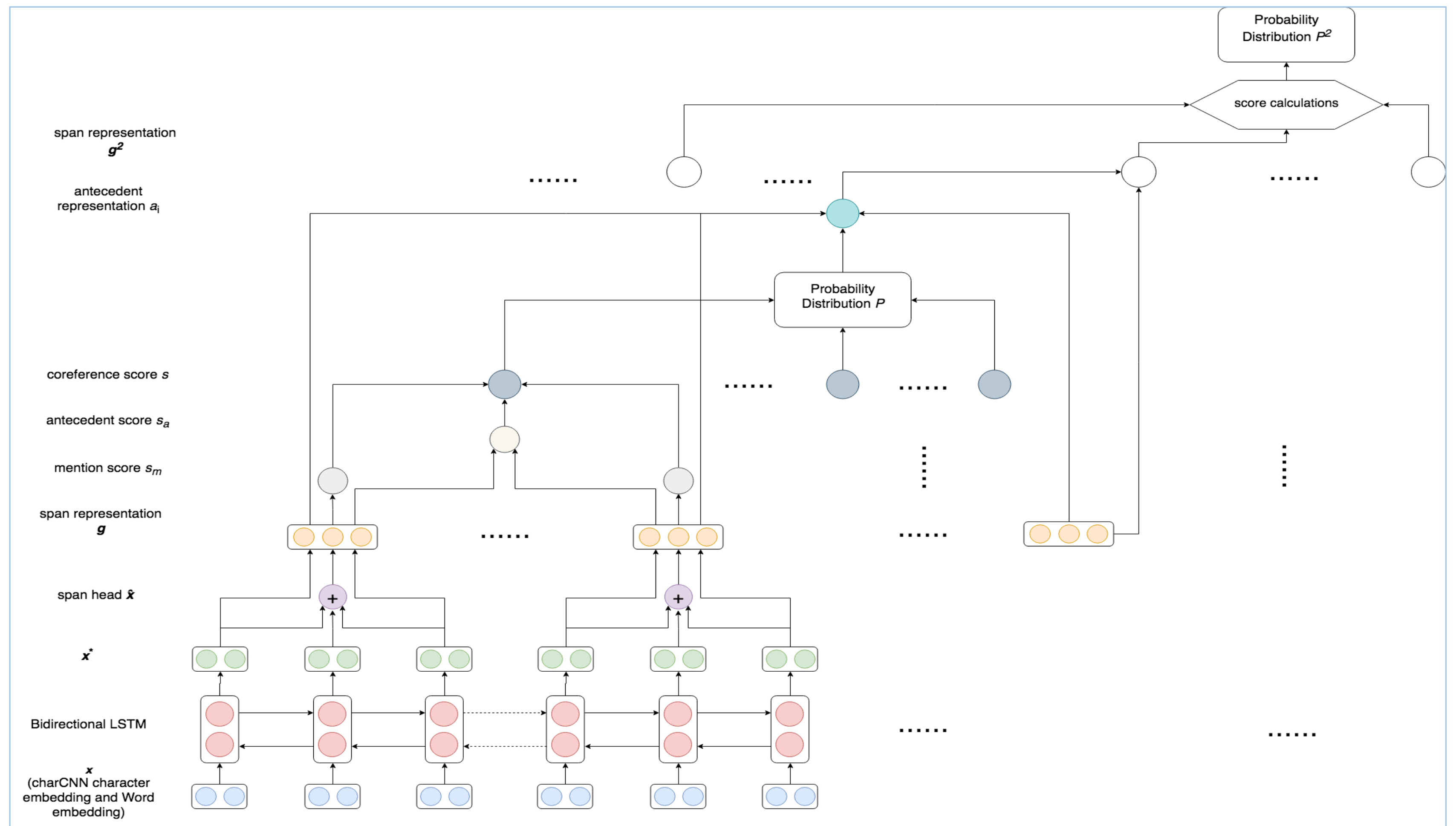
Probability Distribution P^2
score calculations
span representation g^2
antecedent representation a_i
Probability Distribution P
coreference score s
antecedent score s_a
mention score s_m
span representation g
span head $\hat{x}$
x^*
Bidirectional LSTM
x (charCNN character embedding and Word embedding)
......

Figure 1: End-to-end coreference architecture with 2nd order coreferenec resolution - computes embedding representations of spans to predict more likely spans for antecedent matches

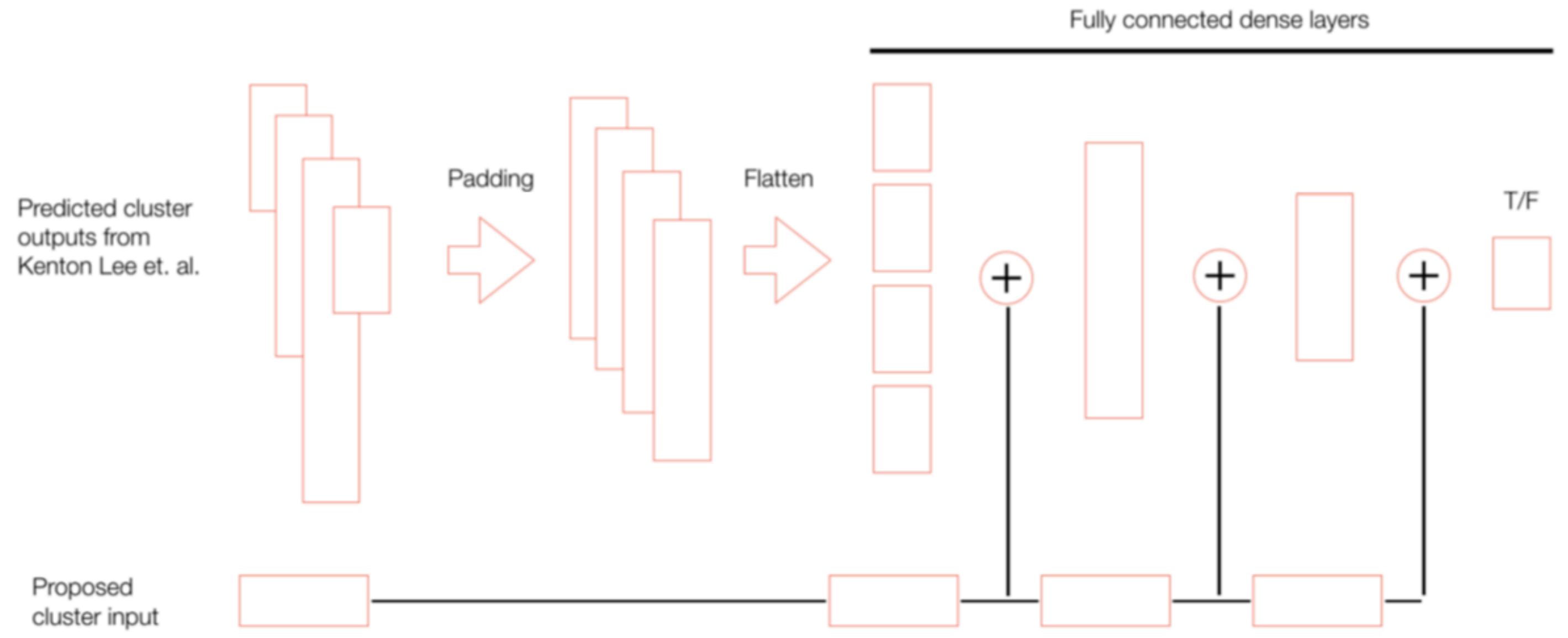
Fully connected dense layers
Predicted cluster outputs from Kenton Lee et. al.
Padding
Flatten
T/F
Proposed cluster input

Figure 2: Residual dense skip-connection layers showing the integration of coreference clusters as input to the model

Attention mechanism + Scoring architecture [1]:

$$\alpha_t = \omega_\alpha \cdot \text{FFNN}_\alpha(\mathbf{x}_t^*) \qquad \mathbf{g}_i = [\mathbf{x}^*_{\text{START}(i)}, \mathbf{x}^*_{\text{END}(i)}, \mathbf{x}_{\hat{(i)}}, \phi(i)] \qquad s_c(i,j) = \mathbf{g}_i^T \mathbf{W}_c \mathbf{g}_j$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=\text{START}(i)}^{\text{END}(i)} \exp \alpha_k} \qquad s_m(i) = w_m \cdot \text{FFNN}_m(\mathbf{g}_i)$$

$$s_a(i,j) = w_a \cdot \text{FFNN}_a([\mathbf{g}_i, \mathbf{g}_j, \mathbf{g}_i \circ \mathbf{g}_j, \phi(i,j)])$$

$$\hat{\mathbf{x}}_i = \sum_{t=\text{START}(i)}^{\text{END}(i)} a_{i,t} \cdot \mathbf{x}_t \qquad s(i,j) = s_m(i) + s_m(j) + s_c(i,j) + s_a(i,j)$$

Second-order Coreference Resolution:

$$P^1(y_i) = \frac{e^{s(\mathbf{g}_i, \mathbf{g}_{y_i})}}{\sum_{y \in Y(i)} e^{s(\mathbf{g}_i, \mathbf{g}_y)}} \qquad \mathbf{g}_i^2 = \mathbf{f}_i \circ \mathbf{g}_i + (\mathbf{1} - \mathbf{f}_i) \circ \mathbf{a}_i$$

$$\mathbf{a}_i = \sum P^1(y_i) \cdot \mathbf{g}_{\mathbf{y_i}}, \mathbf{f}_i = \sigma(\mathbf{W}_f[\mathbf{g}_i, \mathbf{a}_i]) \qquad P(y_i) = P^2(y_i) = \frac{e^{s(\mathbf{g}_i^2, \mathbf{g}_{y_i}^2)}}{\sum_{y \in Y(i)} e^{s(\mathbf{g}_i^2, \mathbf{g}_y^2)}}$$

Metrics

Accuracy was assessed by the F1 score, which is the harmonic mean of precision and recall. Explicitly, the score is calculated as

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where precision is the ratio of true positive guesses to total positive guesses and recall is the ratio of true positive guesses to total ground truth positives. We compute the F1 scores for the standard MUC, B3, CEAFφ4 metrics. The main evaluation metric is the average F1 of the three metrics.

Experiments And Results

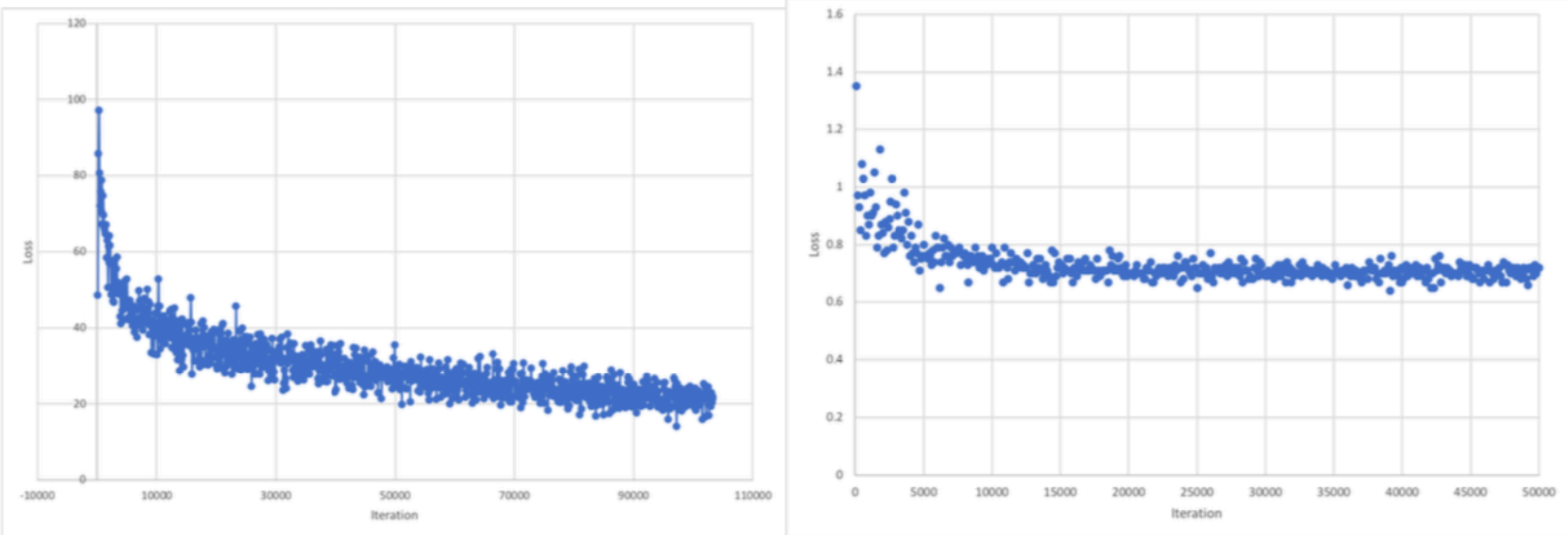

Figure 3: Plot of loss showing model convergence during training. Left - original model using softmax loss, Right - model updated for Google GAP dataset using binary cross entropy loss

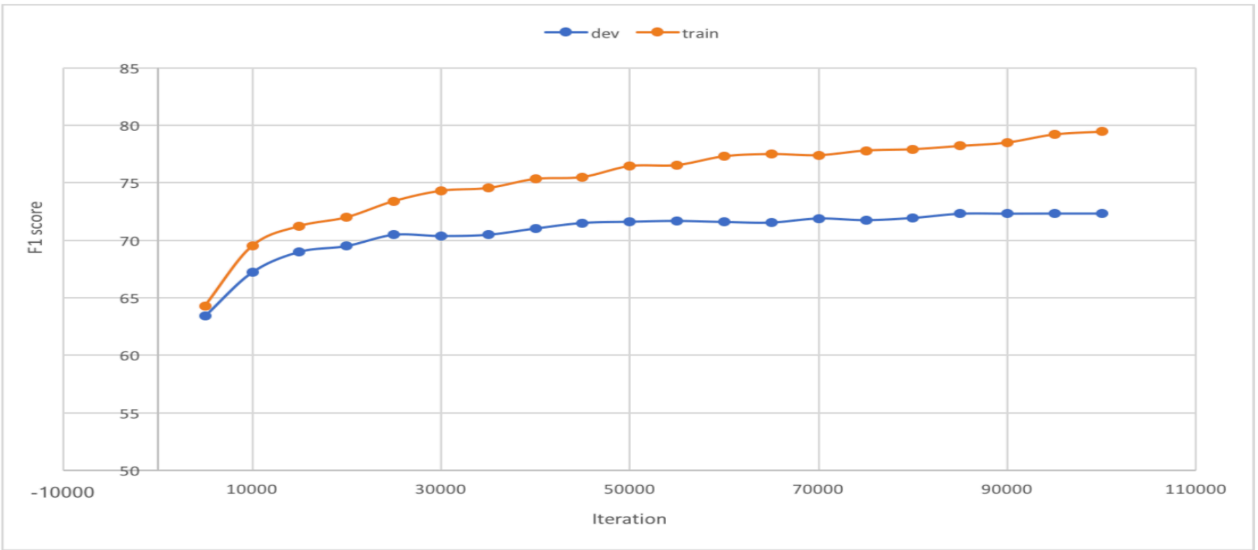

Figure 4: Plot of the F1 performance metric at each epoch, computed on the dev set and the training set, showing model learning during training

	Average F1
Clark and Manning	55.0
Kenton Lee et. al. (2013)	50.5
Kenton Lee et. al. (2017)	64.7
Our updated implementation for GAP	58.2

Table 1: Average F1 scores

Our implementation of the original model achieved an F1 score of 0.7283 on its original dataset. The average precision was 75.76% and the average recall was 70.11%. Our implementation of the updated model achieved an F1 score of 0.5818 on the Google's GAP dataset. The average precision was 58.19% and the average recall was 56.63%.

Analysis

Incorrect coreference resolution example: When onlookers expressed doubt, claiming that the Proctor family was well regarded in the community, the girl promptly came out of her trance and told them it was all for "sport". On March 29, 1692, Abigail Williams and Mercy Lewis again said they were being tormented by Elizabeth's spectre. A few days later, Abigail complained that Elizabeth was pinching her and tearing at her bowels, and said she saw Elizabeth's spectre as well as John's.
Correct coreference: The correct antecedent of 'her' is 'Abigail'.

References

[1] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end Neural Coreference Resolution. arXiv e-prints, page arXiv:1707.07045, Jul 2017.
[2] K. Lee, L. He, and L. Zettlemoyer. Higher-order Coreference Resolution with Coarse-to-fine Inference. arXiv e-prints, page arXiv:1804.05392, Apr 2018.
[3] M.E.Peters,M.Neumann,M.Iyyer,M.Gardner,C.Clark,K.Lee,andL.Zettlemoyer.Deepcontextualized word representations. arXiv e-prints, page arXiv:1802.05365, Feb 2018.
[4] S. Wiseman, A. M. Rush, and S. M. Shieber. Learning global features for coreference resolution. CoRR, abs/1604.03035, 2016.