BERT for Coreference Resolution

Arthi Suresb
artsur@stanford.edu

Abstract

Several downstream NLP tasks including knowledge extraction hinge on effective coreference resolution, the task of determining which noun phrases in text refer to the same real-world entity. In this paper, we focus on a subtask of coreference resolution, ambiguous gendered pronoun resolution (GAP). Noting the ability of Transformer encoders to capture intra-sequence dependencies via self-attention, we explore in this paper the effect of incorporating bidirectional encoder representations from transformers (BERT) into two architectures for coreference - a rule-based heuristic and a mention-ranking model. We also discuss and provide significant code for utilizing BERT in an end-to-end clustering coreference model. Our best model, the mention ranking algorithm using BERT as an embedding layer, achieves an overall F1 of 76.0 and bias of 1.00 on the GAP snippet-context task, improving upon the baseline Parallelism F1 provided in paper by 9.1 and on bias by 0.07.

1 Introduction

We explore in this paper how BERT can be used in a variety of architectures for the GAP coreference resolution task: as (1) an input for a rule-based heuristic, as (2) embeddings in a mention-pair ranking architecture, and as (3) a replacement for a long short-term memory network in a end-to-end neural model that jointly learns mention extraction and coreference clustering.

The coreference resolution task is typically framed as identifying all mentions of entities and events in text and clustering them into equivalence classes [6]. For example, given the sentence, “McFerran’s horse farm was named Glen View. After his death in 1885, John E. Green acquired the farm." an algorithm for coreference resolution would be expected to recognize (“his", “McFerran") and (“horse farm", “Glen View", “the farm") as clusters representing the same real-world entity. Understanding which words are co-referent with one another, as one can imagine, is essential for extracting knowledge, summarizing information, and question-answering tasks. In parts of this paper, we specifically focus on a subtask of coreference resolution, ambiguous gendered pronoun resolution. In ambiguous gendered pronoun resolution, we try to identify the named entity antecedents of gendered pronouns, such as “his" and “her."

While approaches to automating coreference resolution have been discussed as early as 1978 with Hobbs’ naive algorithm, which involves an extensive, complex set of rules over syntactic trees, there have been several machine learning approaches to this problem, the most effective of which, not surprisingly, have been neural methods. However, there are still many challenges with current co-reference systems.

Some practical limitations of current off-the-shelf resolvers is that they are not equally performant on diverse groups, due to the imbalanced representation in training text data. A recent paper by Webster et. al Google AI Language group highlighted that our state-of-the-art resolvers perform considerably better on sentences involving male pronouns as opposed to female pronouns. The then state-of-the-art model by Lee et. al (2017) has a pronoun-antecedent F1 score of 68.9 for male pronouns and 51.9 for female pronouns on the OntoNotes set [6] [7]. Google AI, to motivate building equally effective systems for all genders, released a gender-balanced dataset for gendered pronoun resolution, and

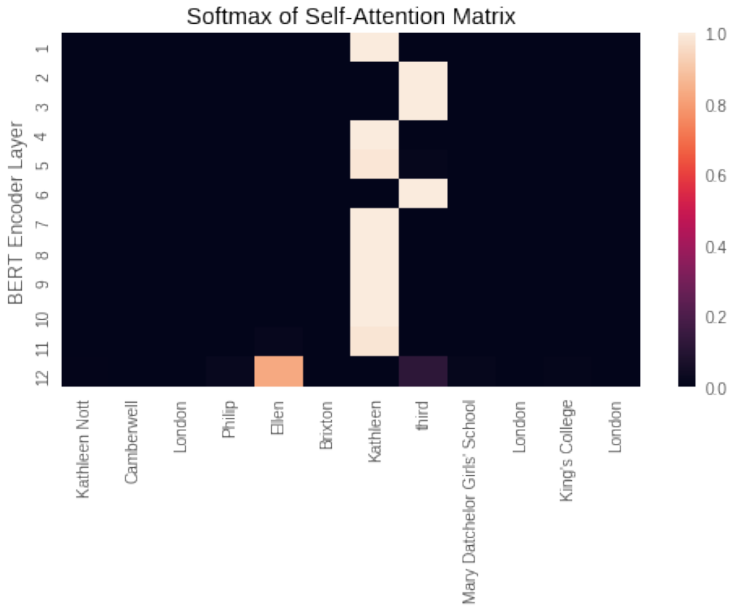

Figure 1: We take as an example a paragraph from the GAP development set, "*Kathleen Nott was born in Camberwell, London. Her father, Philip, was a lithographic printer, and her mother, Ellen, ran a boarding house in Brixton; Kathleen was their third daughter.* ***She*** *was educated at Mary Datchelor Girls' School (now closed), London, before attending King's College, London.*" for which we must identify the antecedent of the bolded "she." We extract the entity mentions in this sentence using spaCy's entity extractor. For each entity, we take the sum of the self-attention matrix values of the ambiguous pronoun "she" to the entity across all 12 attention heads for each encoder layer, and pass each layer's output through the softmax function to produce a probability distribution over the entities. We observe that for this example, 8 of 12 of the encoder layers select "Kathleen" as the correct antecedent.

presented a new NLP task, GAP, that includes as an evaluation metric the bias - the ratio of feminine F1 to masculine F1. For much of this paper, we use this dataset and set of evaluation metrics.

The linguistic challenges in coreference are multifold as well. Many coreference systems do not deal with cataphora (an antecedent appearing after a pronoun) and long-range inferences well. We hypothesize that BERT, bidirectional encoder representations from transformers, learned with the objective of predicting the original vocabulary ID of a randomly masked word based only on its context, can be used to further improve the performance on many of these models, because of the ability of deeply bidirectional self-attention matrices conditioned on both left and right context to capture the kind of intra-sentence dependencies that we need to understand coreferences. The simplest heuristic we can apply on gendered pronoun resolution using BERT is: "select the mention that a pronoun attends to most." In Figure 1, we show a sentence in the GAP development set, and visualize how the ambiguous pronoun "she" attends to the different entities in the encoder attention matrices, in this case, isolating the correct antecedent.

BERT has already been used extensively in question-answering systems and natural language inference tasks, which implicitly need to understand coreferences to extract information, and BERT has been applied to nearly every NLP task with significant improvements to the state-of-the-art [2]. Despite BERT's success in lifting the state-of-the-art on a variety of NLP tasks, not much work has been done to incorporate Transformer models, and BERT specifically, into coreference resolution. This paper tries to scratch the surface of this idea, with promising initial results.

2 Related Work

There are broadly four categories of neural methods for coreference - mention-pair models (in which we make binary coreference decisions on mention pairs), mention-ranking models (in which we

score mentions and make collective decisions), and clustering models (in which we directly predict clusters). Many of these models rely heavily on syntactic parsing for mention or entity extraction algorithms, over which several features are hand-engineered and fed into the model as input. The challenge of these pipelined systems is two-fold: the pre-processing needed to generate features adds complexity, and errors in mention extraction algorithms propagate and limit performance on the final coreference task. In "`End-to-end neural coreference resolution`," Lee et. al present the first architecture that jointly performs the task of generating mention candidates as well as clustering co-references, using only the gold clusters of coreferential mentions for training [4], and word and character embeddings as input. Not only does the Lee et. al architecture simplify the inputs needed for the coreference resolution task, but it also managed to outperform previous state-of-the-art on the OntoNotes set (described later) by 3.1 F1 points. A more recent paper by Lee et. al in 2018 outperforms this model by 5.8 Average F1 on the OntoNotes dataset.

The GAP paper discusses the Lee et. al (2017) model, which achieves an overall F1 of 64.0, as a baseline [7]. The paper also strongly encourages the use of Transformer models for the GAP task, noting that a simple heuristic on a Transformer trained for the English-German NMT task, can produce an F1 score on the GAP dataset of 56.2. We use this as inspiration to apply BERT to coreference resolution.

3 Approach

We explore applying BERT for this task in three different ways: constructing a rule-based heuristic, and applying BERT to different architectures for coreference - a mention-scoring model, and Lee et. al's end-to-end neural coreference model.

Rule-based heuristic: The intuition behind this heuristic is "select the candidate which attends most to the pronoun." We use an entity extractor from spaCy [3] to extract all entities in an input sentence, and utilize BERT self-attention matrices at each layer and head and choose the entity that is attended to most by the pronoun. A similar approach is used in the GAP paper with the Vaswani et. al Transformer model. Since we use WordPiece tokenization, we calculate the attention between two subtokens as the sum of the attention between the pronoun and all occurrences of the subtoken; and the attention between the pronoun and the candidate antecedent as the mean of the attention between the pronoun and all subtokens in the candidate antecedent. We explore several methods of combining these 12 layers and 12 heads to produce a final decision on the GAP dataset.

Mention-scoring approach: The second architecture takes in as input a passage, extracts potential antecedents using spaCy entity extractor, and constructs several hand-assembled features about a pronoun-candidate entity pair which are described in Clark et. al, and described further in section 3. The input representing each pronoun-entity pair is passed into a feed-forward neural network with three hidden layers of rectified linear units, and projected onto an output layer to produce a "pronoun-entity pair score"

We describe the entity-scoring model, heavily inspired by Clark and Manning (2017) [1], in Figure 2. The score for the dummy antecedent, ϵ representing the case where there is no antecedent in the passage, or it is unknown, is always set to 0. To obtain a probability distribution, we apply a softmax over the scores all of the entities extracted from the sentence to produce $\Pr(p_i, e_i)$ for each potential antecedent for pronoun p_i, and pass this into the loss function. The model is trained to minimize the negative log-likelihood of the correct antecedent for the pronoun, among all of the extracted entities. In the case where neither of the two provided options are the antecedent, we mark as the correct label, the dummy antecedent, ϵ. The batched loss function can be formalized as such:

$$\text{J} = \sum_{p_i \in (\texttt{minibatch})} \left(-\log \left(\sum_{e_i \in (E \cup \epsilon)} \mathbb{1}(y_{p_i,e_i}) \Pr(p_i, e_i) \right) \right)$$

where E is the set of named entities $e_1, ..., e_n$ extracted from the text, p_i is a pronoun in the minibatch, and $\Pr(p_i, e_i)$ is determined by the softmaxed entity-pronoun scores.

We add BERT to this model by taking the last layer's self-attention outputs as input embeddings rather than using GloVe embeddings as input to this model. We implement this model from scratch.

End-to-end coreference approach: The last architecture we explored is an end-to-end neural coreference model that jointly performs mention extraction and coreference clustering, by optimizing

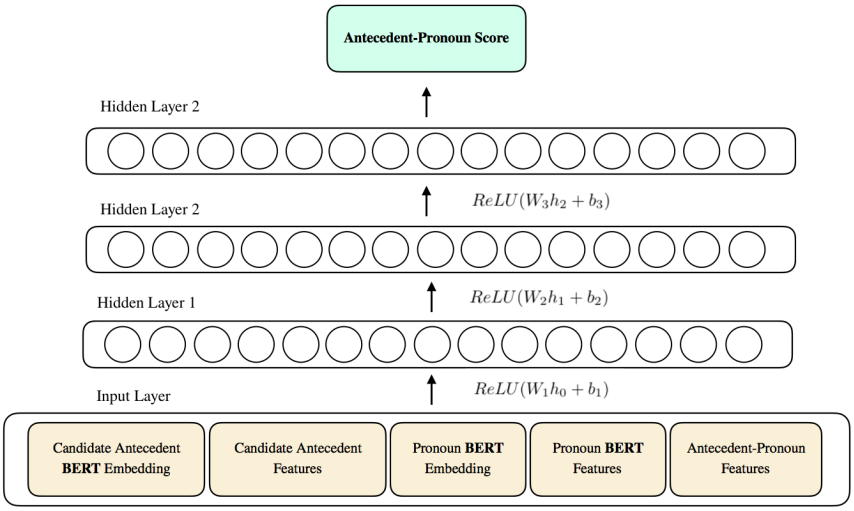

Figure 2: Pronoun-Candidate Antecedent Scorer

the marginal log-likelihood of all correct antecedents implied by the gold coreference clustering. We train this baseline model on the OntoNotes dataset [6]. While add BERT to this model by replacing the bidirectional LSTM component of the model whose purpose, as described in Lee et. al (2017), is to "encode each word in its context" with averaged last four encoder outputs. For further information, we refer the reader to the original paper [4].

We use as our baselines, the Lee et. al model and the Parallelism syntactic heuristic, whose performance on the GAP test set is reported in Table 2.

4 Experiments

4.1 Data

4.1.1 Source

For the mention-ranking model, we use the gender-balanced GAP dataset published by Google AI, released to motivate building equally effective coreference resolution systems for masculine and feminine genders [7]. This dataset, generated using text from Wikipedia, consists of a collection of 2000 training examples, 2000 test examples, and 454 validation examples, consisting of two mention-pronoun labels each. A single example in the dataset contains a passage, the location of an ambiguous pronoun, the location of two potential antecedents of the pronoun, `A` and `B`, and a label of whether the pronoun refers to `A`, `B`, or `Neither`.

For experiments on the end-to-end coreference model, we use the English OntoNotes corpus, consisting of over 1 million words from newswire, magazine articles, broadcast news, broadcast conversations, web data, telephone conversations, and English translation of the New Testament. There are 2802 training documents, 343 development documents, and 348 test documents in this corpus. We use this dataset since it contains comprehensive information on gold clusters necessary for the loss function, rather than on two mention-antecedent pairs.

4.1.2 Representation

For our mention-ranking model baseline we use as input fixed 300-dimensional GloVe vectors from spaCy [3] to represent the first word and last word of each candidate antecedent, the two preceding words, the two following words, the head word of the mention, the dependency parent (determined using the spaCy dependency parser), the average of the five following words, the average of the five preceding words, and the average of the sentence. We additionally encode information regarding

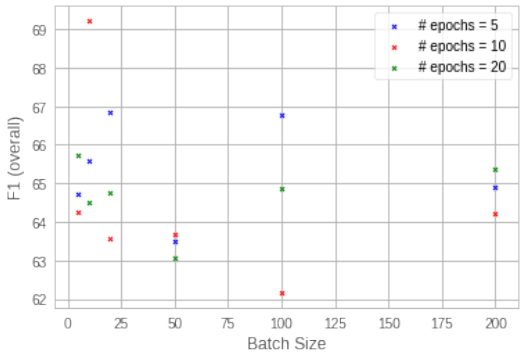

Figure 3: Example of hyperparameter search with learning rate 0.001.

Table 1: Heuristics on BERT Attention Matrices

Heuristic	Overall GAP F1
Mode of highest-scoring entities across layers	40.4
Highest-scoring entity, averaged last 4 layers	40.4
Highest-scoring entity, last layer	27.9
9th layer, 11th head	68.6

the distance between the pronoun and each of the mentions in terms of tokens and characters and length of the mention in the continuous and vectorized forms suggested in Clark, et. al [1], and whether the entity is contained within another entity. The input representation for each example is a concatenation of all of the aforementioned features, and its dimensionality is $k = 7546$.

We add BERT by replacing 300-dimensional GloVe vectors with 768-dimensional vectors from the pretrained `bert-base-uncased` BERT encoder output, resulting in 18478-dimensional vectors for each (`pronoun`, `named entity`) pair.

4.2 Evaluation Method

The GAP task is evaluated using four metrics: F1 score (harmonic average of the precision and recall) on masculine (M) and feminine (F) examples, overall F1 score (O), and a bias factor (B) calculated as the ratio between feminine F1 and masculine F1, where the bias factor in most state-of-the-art production systems is less than 1.

4.3 Experimental Details

For the mention ranking model, use Adam optimizer with the aforementioned loss function and a learning rate of 0.001, and train for 10 epochs with batch size of 100. The first hidden layer has 1000 hidden units, and the second two layers have 500. We use dropout after each hidden layer with probability 0.6 to prevent overfitting the small training dataset. We performed mild hyperparameter tuning on the validation set to select an appropriate combination of (batch size, number of epochs, learning rate), as shown in Figure 3.

For our rule-based heuristic, we explored different ways of combining the attention matrices, which are detailed in Table 1. The most effective rule we encountered was using a single attention head and layer.

In addition, we spent a significant amount of time reworking an implementation of the end-to-end neural coreference model in PyTorch, with code heavily borrowed from `shayneobrien`'s coreference-resolution repository, and attempted to fix and run the model with all of the same hyperparameters and architectural choices laid forth in the Lee et. al paper [5]. The model, run for 110 epochs instead of 150, produced an MUC about 24.5 F1 below the expected 77.2 laid forth in the paper. We have

Table 2: Our models and baselines on the GAP challenge test set

Name	M	F	B	O
Lee et. al (2017)	67.7	60.0	0.89	64.0
Parallelism	69.4	64.4	0.93	66.9
BERT Attention Heuristic	71.2	65.9	0.93	68.6
Mention-Scoring Model + GloVe	66.3	63.0	0.95	64.7
Mention-Scoring Model + BERT	**75.9**	**76.0**	**1.00**	**76.0**

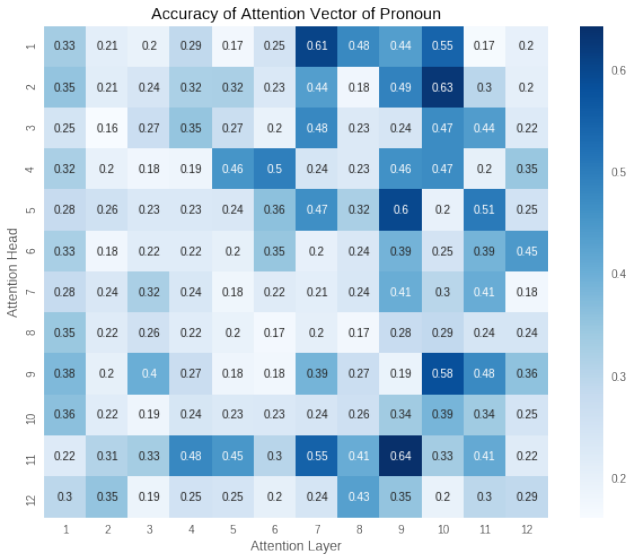

Figure 4: Classification accuracy of BERT self-attention matrices.

coded a skeleton for how to incorporate BERT as a Document Encoder layer, but leave it as future work to rectify and report the findings of this approach.

We also note that we explored a few different approaches to the GAP problem not detailed here. Initially, we treated this as a multiclass classification problem, as does the "Gendered Pronoun Resolution" Kaggle challenge. We veered from this track as it does not represent the GAP challenge as posed, in which gold mentions must be inferred rather than used directly in the input layer to a network, and represents a less practical scenario of having only three predefined options for an antecedent.

4.4 Results

We achieve an improvement of 9.1 overall F1, 6.5 masculine F1, 11.1 feminine F1, and 0.07 absolute on bias with our mention ranking algorithm which utilizes BERT as embeddings.

We surprisingly achieve an improvement over the baselines set in the GAP paper of 1.7 overall F1, 1.8 masculine F1, and 1.5 feminine F1, by just using a single attention matrix from a pretrained BERT model, corresponding to the 11th head and 9th layer encoder unit.

5 Analysis

Given the lift BERT has provided to presumably more difficult NLP tasks such as question answering and translation, it makes sense that BERT would be able to improve performance on coreference resolution as well. In Figure 4, we visualize how each of 144 attention matrices (12 layers, 12 heads) in the BERT model perform on the GAP validation set. We can see that deeper encoder units have a

stronger coreference signal, and that the signal seems to be localized to specific attention heads. The GAP paper notes this too [7].

It is even more promising that the two highest-accuracy attention heads seem to get different examples correct, as show in Figure 5, which demonstrates that these likely capture different aspects of coreference, and can improve effectiveness of the heuristic approach when used together, and improve effectiveness of models when used as input features.

There are still some limitations with BERT. We show that the accuracy of the model that uses BERT degrades as the distance between the pronoun and correct antecedent increases, observing only the cases where the correct antecedent is given mention A or mention B. This may highlight that BERT still suffers.

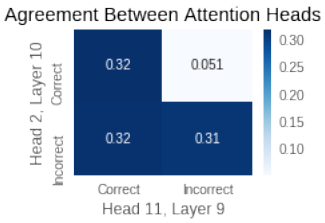

Figure 5: Agreement of highest-accuracy BERT attention heads on validation set.

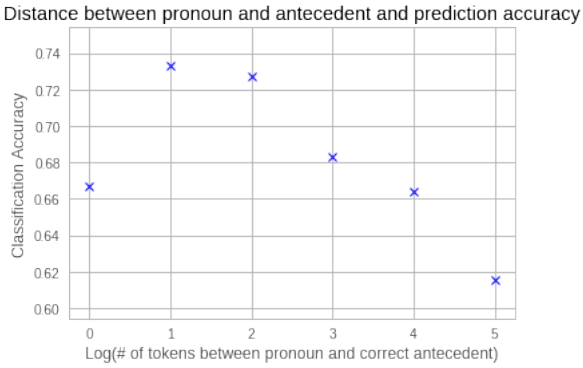

Figure 6: Classification accuracy by distance between pronoun and gold antecedent.

Another limitation of this model is the dependence on the named entity extraction step. We observed that of the 88.65% of examples in the GAP test set that have A or B, the entity extraction model from spaCy did not correctly extract the correct antecedent in 7.27% of cases. This hugely reduces the upper bound of model performance, and validates the risks of using a pipelined system.

6 Conclusion

From these preliminary results, we conclude that BERT can be applied successfully to improve the performance on the GAP coreference resolution task. We observe that BERT attention matrices are likely able to do this by effectively encoding the coreference signal in deeper layers, and at specific heads.

6.1 Future Work

There are many avenues for future work. As described, there was extensive effort in implementing and training the end-to-end neural coreference Pytorch implementation, but we were not able to match the performance in the Lee et. al paper and so we leave it as future work to implement BERT as a replacement for the LSTM component in this model. Much of the code for this modification of Lee et. al's model is already written, and available on Github. Here, we heavily used and modified the repository provided by shayneobrien [5]. The limitations of our current model, as shown in the previous section, include that it is dependent on successful entity extraction. With and end-to-end model that jointly extracts mentions, we are bound less by this constraint.

One challenge of the models presented in this paper is that the uses BERT simply as a pre-trained embedding, and cannot reap any of the further lifts in performance from fine-tuning BERT for the end-to-end task. We found it difficult to frame the coreference task as one of the tasks for which a BERT model for fine-tuning is available, and future work can include effectively doing so.

7 Additional Information

My mentor for this project was Xiaoxue Zang. For further reference, please refer to the GitHub repositories linked below for the aforementioned code:

`https://github.com/arthisuresh/gap-coreference`
and
`https://github.com/arthisuresh/coreference-resolution`

References

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

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

[4] Kenton Lee, Luheng He, Mike Lewis, and Luke S. Zettlemoyer. End-to-end neural coreference resolution. In *EMNLP*, 2017.

[5] Shayne O'Brien. Efficient and clean pytorch reimplementation of "end-to-end neural coreference resolution" (lee et al., emnlp 2017). `https://github.com/shayneobrien/coreference-resolution`, 2017.

[6] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea, 2012.

[7] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered ambiguou. In *Transactions of the ACL*, page to appear, 2018.