BERT for Coreference Resolution

Arthi Suresh (artsur@stanford.edu)

Problem

- **Coreference resolution** is the task of identifying clusters of mentions referring to the same real-world entity. Given the sentence "***She had a good suggestion and it was unanimously accepted***" a coreference resolver would be expected to recognize ***it*** and ***good suggestion*** as coreferrent mentions.
- **BERT**, bidirectional encoder representations from transformers, has lifted the state-of-the-art in a variety of NLP tasks but have not been applied directly to coreference resolution, a task in which self-attention that can capture intra-sentence dependencies could be immensely useful.

Task & Data

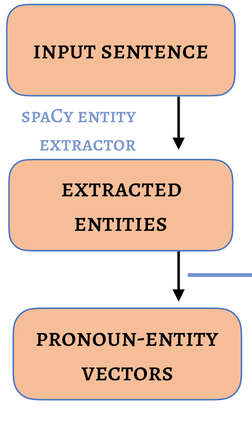
GAP CHALLENGE

- We focus in this paper specifically on the task of **gendered pronoun resolution**, as presented in the GAP dataset [1]. task of identifying for a specified pronoun in a passage, which named entity antecedent it refers to.

Sentence	Pronoun	A	B
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.	she	Kathleen (TRUE)	Ellen (FALSE)

- A model is evaluated on how well it predicts the outputs (A, B, or NEITHER) using F1 score for all examples (O), masculine examples (M), feminine examples (F), and the ratio between F/M, called the bias (B).

DATA PROCESSING

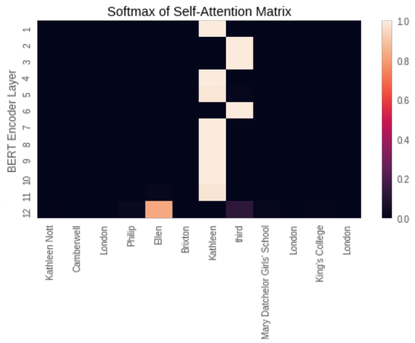


1. Compute several embedding features about the candidate antecedent and the pronoun-entity pair, such as distance (refer to [2] for further details).
2. When using **BERT** for the attention heuristic, we extract the row corresponding to the pronoun in the attention matrices of the encoder units. When using **BERT** for the vector representations into the mention scoring model, we take the final encoder output from each layer, using the first subtoken to represent multi-subtoken entities.

Approach

We study two methods of using BERT for coreference resolution.

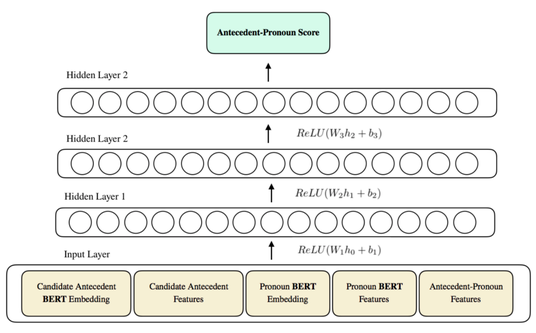
ATTENTION HEURISTIC



- Using the row corresponding to the pronoun in the l x l x 768 attention matrices from each encoder unit from 12 heads x 12 layers, we use the simple heuristic of "**selecting the entity the pronoun most attends to.**"
- We tried many methods of combining these 144 values, but found that simply using a single attention head and layer was effective.

MENTION SCORING MODEL

- We pass the vectors representing each pronoun-entity pair through a FFNN with three hidden layers with 1000, 500, and 500 hidden units, respectively.



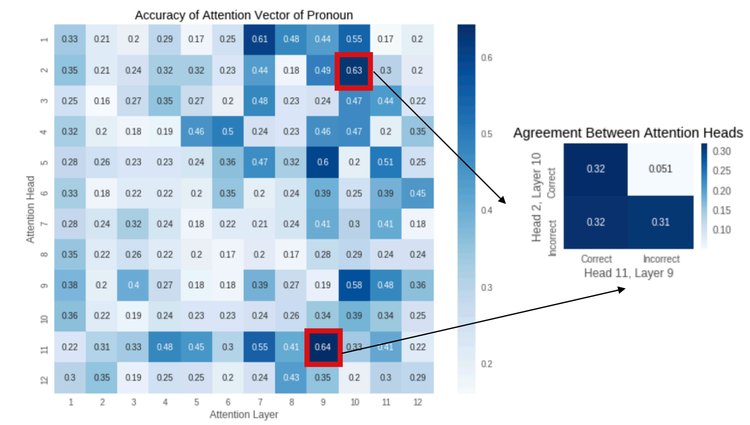
- We take the softmax of the scores for all of the extracted entities in a sentence, and a dummy antecedent whose score is set to 0 to obtain a probability distribution over the candidates.
- We minimize the **negative log-likelihood of the correct antecedent for the pronoun**, among all of the extracted entities.

Results

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**
	+6.5	**+12.0**	**+0.07**	**+9.1**

- Our best architecture, a ***mention-scoring model that uses BERT as an input embedding layer***, beats the baselines laid forth on the snippet-context task by 9.1 overall F1, 6.5 masculine F1, 12.0 feminine F1, and 0.07 on bias.

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.
- We visualize how each of 144 attention matrices 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**. Even more promising, we see that the two highest-accuracy heads get different examples correct, demonstrating promise in using these as input features.
- A limitation of this model is the dependence on the named entity extraction step. Of 88.65% of examples in the GAP test set that have A or B as the true antecedent, the entity extractor did not correctly extract the correct antecedent 7.27% of the time, upper bounding the performance of the end-to-end model.

Conclusions & Future Work

- BERT can be applied successfully to improve the performance on the GAP coreference resolution task.
- To avoid limitations of a pipelined system, future work includes incorporating BERT into the end-to-end joint mention extraction and coreference clustering model presented by Lee et. al (2017) as a replacement to the LSTM whose function it is to "encode each span within its context," a job we hypothesize BERT does well.

References

[1] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered ambiguous pronouns. In *Transactions of the ACL*, 2018.

[2] 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), 2016.