Interpreting BERT for downstream tasks: Empirical Study on Coreference Resolution

Parth Shah
University of California, Los Angeles
parthushah8@g.ucla.edu

Jing Xu
University of California, Los Angeles
xujing98@g.ucla.edu

Jieyu Zhao
University of California, Los Angeles
jieyuzhao@ucla.edu

Fan Yin
University of California, Los Angeles
fanyin20@g.ucla.edu

Abstract

Since the introduction of BERT, numerous benchmarks have been created by fine-tuning pre-trained language models. Although interpreting results of a BERT-based model has been a pressing concern, as we use them to solve critical real-life tasks. Many attribution techniques have been suggested to explain BERT models, but they are often limited to explaining sequence encodings. Our study adapts existing methods of attribution to explain the decision-making of BERT on the downstream task of coreference resolution. We employ two attribution methods by applying them to the state-of-the-art BERT-based coreference resolution model. This work also provides guidance on using attribution methods to better understand how BERT makes decisions for downstream tasks.

1 Introduction

There has been an upsurge in the development of fine-tuned models that outperform existing benchmarks on various NLP tasks since the introduction of pre-trained language models, such as ELMO (Peters et al. [2018]) and BERT (Devlin et al. [2018]). Furthermore, advanced BERT variants have been developed since the debut of BERT, such as RoBERTa (Liu et al. [2019]), ELECTRA (Clark et al. [2020]), and XLNet (Yang et al. [2019]) which have achieved state-of-the-art results on various NLP tasks. The effectiveness of such models has prompted curiosity in how BERT was able to perform at a human level on various tasks. Such research is necessary in order to improve the transparency of deep neural networks, an imperative property for critical NLP tasks having high stakes. It further benefits in comprehending feature importance for real-world problems, which may aid in drawing critical conclusions.

A variety of attempts have recently been made to explain BERT. Work by Tenney et al. [2019] and Clark et al. [2019] analyzed local attention weights, suggesting that they carry syntactic and semantic information. Parallelly various attribution methods, such as analyzing head importance and probing structural properties learned for sequence-to-sequence tasks, were adapted to explain BERT (Voita et al. [2019]). However, such attribution methods are limited to interpreting models consisting solely of BERT and need to be reworked for applying those to any downstream tasks using BERT. In this work, we try to bridge that gap by interpreting the results of the BERT c2f-coref model introduced by Joshi et al. [2019] used for the coreference resolution task which uses a BERT encoder in the coref model instead of LSTM originally used by Lee et al. [2018].

The BERT-based c2f-coref model consists of multiple feed-forward neural networks on top of the embedding to identify spans and their corresponding cluster using other meta-data as features, BERT being a crucial but not the sole component of the model. We use two attribution methods to interpret our results, comparing attention matrices with their importance scores as per Clark et al. [2019] and using Layer Integrated Gradients introduced by Sundararajan et al. [2017] Through this work, we aim to generalize interpreting models using neural language encoders to solve the final downstream task. Also, our study focuses on the Winobias dataset developed by Zhao et al. [2018], comparing interpretation of sentences having different gendered pronouns to gather insights about the model's shortcomings.

2 Coreference Resolution

Coreference resolution is the task of finding all expressions that refer to the same entity in a text. It's a key step in a variety of higher-level NLP tasks including document summarising, question answering, and information extraction that need natural language understanding. We use the end-to-end c2f-coref model introduced by Lee et al. [2018] replacing the original LSTM-based encoder to the BERT-based encoder which gives superior performance in the *independent* setup as suggested by Joshi et al. [2019].

2.1 End-to-End Coreference System

The input to the end-to-end c2f-coref model is a text $\mathcal{T}$ containing n words along with it's meta-data containing the speaker and genre information. Given those n words there are $\frac{n(n+1)}{2}$ number of possible spans denoted by their starting and ending indices $start_i$ and end_i, $0 \leq i < \frac{n(n+1)}{2}$ and $0 \leq start_i \leq end_i < n$, ordered based on $start_i$; spans starting at the same index are then sorted based on end_i. For each candidate span x_i the c2f-coref model learns the distribution over its antecedents $y \in \mathrm{Y(x}_i) = \{\epsilon, x_0, \ldots, x_{i-1}\}$, a dummy antecedent ϵ and all preceding spans. The dummy antecedent represents the two possible scenarios, span not being an entity mention or an entity mention that is not coreferent with any previous spans.

For each candidate span x, the model learns the distribution over its antecedents $y \in \mathcal{Y}(x)$ is given by:

$$P(y) = \frac{e^{s(x,y)}}{\sum_{y' \in \mathcal{Y}(x)} e^{s(x,y')}}$$

where $s(x, y)$ is the local score involving two parts: how likely the spans x and y are valid mentions, and how likely they refer to the same entity:

$$\begin{aligned} s(x,y) &= s_m(x) + s_m(y) + s_c(x,y) \\ s_m(x) &= w_m \mathrm{FFNN}_m(g_x) \\ s_m(y) &= w_m \mathrm{FFNN}_m(g_y) \\ s_c(x,y) &= w_c \mathrm{FFNN}_c(g_x, g_y, \phi(x,y)) \end{aligned}$$

g_x, g_y are the span embeddings of x and y, $\phi(x, y)$ is the meta-information (e.g., speakers, distance), and w_m, w_c are the mention and coreference weight matrices, respectively (FFNN: feedforward neural network).

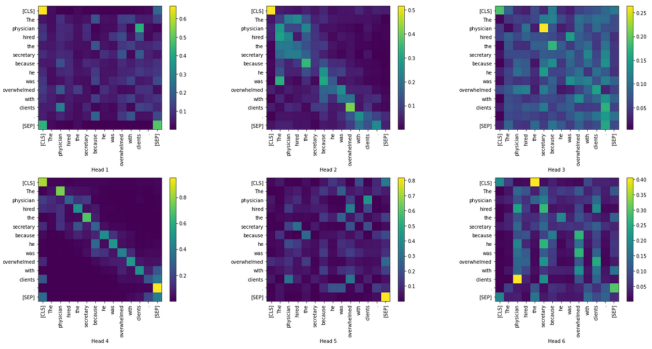

Figure 1: Token-to-Token Head Attention scores

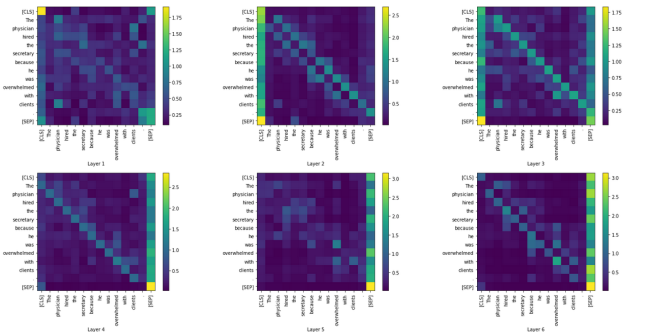

Figure 2: Token-to-Token Layer Attention scores

2.2 Applying BERT

The original model used LSTM-based encoding (with ELMo and GloVe embeddings as input) which has since been replaced with a BERT transformer by Joshi et al. [2019]. A span representation is given by the first and last word pieces, concatenated with the attended versions of all word segments. As BERT is trained on at most 512-word segments, each document is split into segments of maximum segment length. Joshi et al. [2019] tried with two variants of splitting the document: *independent* and *overlap*. In the *independent* variant, each segment acts as a separate BERT instance. In this case, each token is represented by the set of words that lie within its segment. The *overlap* variant was introduced in order to address the limited encoding capability of the *independent* variant. A $\mathcal{S}$-sized segment is created after every $\mathcal{S}/2$ tokens in this variant of document splitting to create overlapping segments. Our experiments, however, used the *independent* variant because Xu and Choi [2020] found it to be more effective on the basis of empirical evidence.

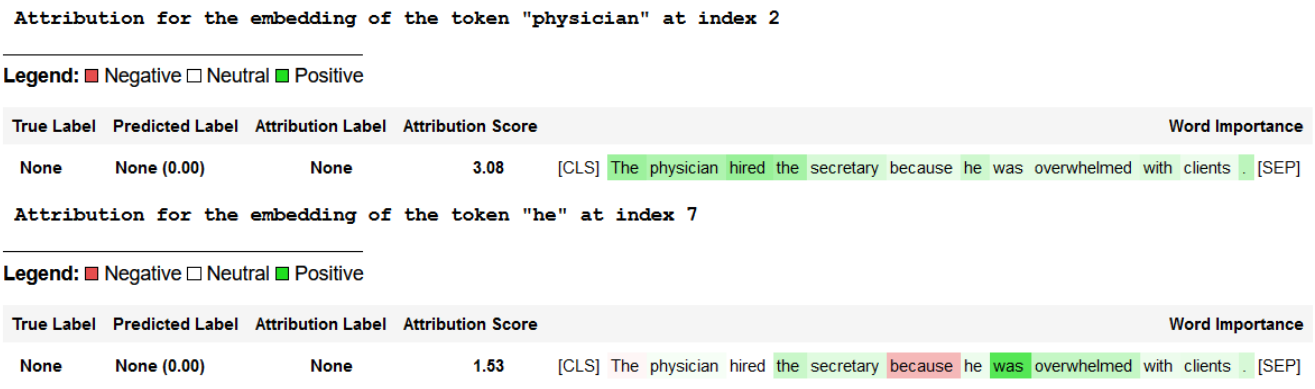

Figure 3: Head Attention scores

3 Attribution Methods

3.1 Attentions Matrices

Among the main characteristics of BERT Transformers is that they utilize a mechanism of attention without using recurrent or conventional layers. Based on a sequence of input vectors, an output vector is calculated by accumulating relevant information through the attention mechanism. Specifically, it assigns attention weights (i.e., relevance) to each input vector and then sums the weighted input vectors to get the final output. Previous research has examined the relationship between attention weights and various linguistic phenomena (Clark et al. [2019], Kovaleva et al. [2019], Reif et al. [2019], Lin et al. [2019], Htut et al. [2019]).

Therefore, we examined the attention matrices, also known as attention probabilities, of each of the 12 layers and each of the 12 heads of the BERT transformer in order to gain a better understanding of the results. The softmax-normalized dot-product represents the relationship between the key and query vectors. Literature in the past, Clark et al. [2019], has shown this to be an indication of how much a token attends/relates to another token in the text. As an example, when it comes to translation it provides a good indication of how much attention has been paid to a token in one language and its corresponding translation in another. When it comes to the Question Answering model, it indicates which tokens are related to each other within the question, text, or answer segments.

Figure 1 shows the attention scores for an example from the Winobias dataset, *The physician hired the secretary because he was overwhelmed with clients*, for the first six heads of the first layer. Based on the visualizations above we observe that there is high attention set along the diagonals. This, however, does not provide us with a significant means to identify other tokens that will have a significant effect on the result collectively. When we examine another layer, we observe a pattern similar to the one above.

We computed and visualized the L_2 norm across all the heads for all the layers in the BERT transformer in Figure 2, which gives us a summary of each layer across all the heads. These visualizations, however, do not convince us that attention scores are a reliable measure of importance for token-to-token relationships across all layers. There are some instances when we observe a strong signal along the diagonal and for the `[SEP]` and `[CLS]` tokens. In this respect, these signals are not true indicators of what semantic the model is learning. Consequently, in order to properly interpret the model, we investigated another popular approach.

3.2 Layer Integrated Gradients

The integrated gradients introduced by Sundararajan et al. [2017] is an axiomatic model interpretability algorithm that assigns an importance score to each input feature by calculating the integral of gradients of the model's output with respect to the inputs along the path (straight line), starting at given baselines/references to the inputs and to approximate the integral we use the Gauss-Legendre quadrature rule. Layer Integrated Gradients is a variant of Integrated Gradients that assign importance scores to layer inputs or outputs based on whether they are attributed to inputs or outputs.

As a baseline, we replaced all text tokens with PAD tokens. Below are the attributions of a few of the words embedded by the BERT transformer

using this approach. Using Layer Integrated Gradients we were able to compute the attributions with respect to the BERT embeddings of tokens, which can be seen in Figure 3. Based on the results, we can tell that the token embeddings are highly dependent on their nearby tokens, and we also observed that in the coreference resolution examples, the span tokens were highly dependent on those of other span tokens. This is a validating observation for the new attribution technique. In addition to the token embeddings, we are also exploring the possibility of interpreting the final result using the Layer Integrated Gradient technique instead of just limiting it to BERT token embeddings, which we believe will provide more interesting insights.

4 Future Work

In the future, we plan to use other attribution methods encompassing the interpretation of the complete forward pass through the model. This involves interpreting each step: identifying probable spans and scoring each span pair based on them identifying the same object at a single step. Furthermore, we plan to generalize and document our method so that it can be applied to interpret any BERT-based model that uses the encoding from the transformer for downstream analyses so that the interpretation of all such analyses is enabled. Eventually, we intend to utilize these interpretations to guide our model in its learning process, possibly resulting in a faster and more stable learning process and also increasing its generalizability.

References

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341, 2019. URL http://arxiv.org/abs/1906.04341.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555, 2020. URL https://arxiv.org/abs/2003.10555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. Do attention heads in BERT track syntactic dependencies? *CoRR*, abs/1911.12246, 2019. URL http://arxiv.org/abs/1911.12246.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. BERT for coreference resolution: Baselines and analysis. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1445. URL https://aclanthology.org/D19-1445.

Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. *CoRR*, abs/1804.05392, 2018. URL http://arxiv.org/abs/1804.05392.

Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4825. URL https://aclanthology.org/W19-4825.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL http://arxiv.org/abs/1802.05365.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. URL http://arxiv.org/abs/1703.01365.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. *CoRR*, abs/1905.05950, 2019. URL http://arxiv.org/abs/1905.05950.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580.

Liyan Xu and Jinho D. Choi. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533. Association for Computational Linguistics, November 2020. URL https://www.aclweb.org/anthology/2020.emnlp-main.686.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019. URL http://arxiv.org/abs/1906.08237.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876, 2018. URL http://arxiv.org/abs/1804.06876.