Some linguistic correlates of gradients and attention weights in BERT

Matthijs Westera
Universitat Pompeu Fabra
Barcelona, Spain
matthijs.westera@upf.edu

Abstract

This work applies the well-known BERT model to a selection of part-of-speech tagged, dependency-parsed and coreference-annotated text, extracting gradients and attention weights for inspection. This reveals that, in BERT, more information flows from a noun to a pronoun if they corefer; open-class words are generally more informative than closed-class words; and there is a slightly underwhelming correlation between BERT's gradients and dependency parses. It also highlights that attention weights and gradients are of course correlated, but they do not always reveal exactly the same patterns.

1 Introduction

Recent years have seen the advance of powerful, general language models as foundational building blocks of more specialized models. The *contextualized* token embeddings of these models contain information that is useful for many downstream NLP tasks. But this information doesn't come out of nowhere; the model must derive it somehow from the context-invariant token embeddings at the start (or, depending on the model, character or word-piece embeddings). This suggests that one can try to understand what a model is doing by investigating not (just) the type of information contained in the contextualized word embeddings, but by considering how it got there. This is not a new idea, of course, but most approaches (e.g., Hewitt and Manning 2019) do concentrate foremost on the representations. The work reported here investigates to what extent one can correlate information flow with certain linguistic phenomena. This is done using the pre-trained transformer model BERT for English (Devlin et al., 2018), although most of the methodology generalizes to other kinds of models.

2 Method

BERT is a transformer, in a nutshell, a stack of self-attention layers. For practical reasons I use the basic version of BERT in this study, which has 12 layers of 12 attention heads each. With regard to a transformer model like BERT, one can consider measuring 'information flow' from a token t_1 to a token t_2 in any number of ways. In this work I concentrate on the following three methods: $G_n(t_1, t_2)$ denotes the magnitude (2-norm) of the gradient of the token embedding of t_2 at layer n with respect to the token embedding t_1 at the previous layer $n-1$. $G^*_n(t_1, t_2)$ denotes the same but with respect to the embedding of token t_1 prior to any of the self-attention layers. $A_n(t_1, t_2)$ denotes the mean attention weight (averaged over all attention heads of a layer) of token t_2 at layer $n-1$ with respect to token t_1 at layer $n-1$ (hence, how much t_2 at layer n depends on it).

Attention weights and gradients are correlated of course: the more attention t_2 pays to t_1, the larger its gradient with respect to it. But across tokens no perfect correlation is expected, e.g., the token to which the most attention is paid is not necessarily the token with respect to which the gradients are the largest. Indeed, Pearson correlation coefficients between attention weights and gradients, on data described below, range between 0.19 and 0.43 (mean 0.37; p-values around 0.001).

3 Data

I consider three linguistic domains where one might expect information flow to show certain patterns: coreference, parts of speech, and dependency trees. Concerning **coreference**, my hypothesis is that more information flows from noun to coreferring pronoun than to non-coreferring pronoun. To test this hypothesis I take 500 random sentences from coreference-annotated OntoNotes

(Weischedel et al., 2013) (development set), each containing a noun followed by (at some distance) a pronoun, co-referring in half of the sentences and not in the other. This lets one compare the amount of information flow from noun to pronoun in the two conditions.

For **parts of speech**, I take 500 POS-annotated random sentences from the GUM portion of the Universal Dependencies dataset for English (Zeldes 2017; development set). My hypothesis is that open-class words (such as nouns, adjectives, verbs) tend to provide more information than closed-class words (such as quantifiers, connectives). I use the Universal Dependencies classification of 'open' vs. 'closed'; I also compare the main parts of speech (noun, verb, adjective, etc.) independently of the open/closed distinction.

For **dependency trees**, I use the same GUM portion of the Universal Dependencies dataset, but this time use its dependency tree annotations. My hypothesis is that information flows primarily along dependency branches, e.g., more from verb to object and less from subject to object. To test this I compute the Pearson correlation of the attention/gradient matrix and the distance matrix of the gold dependency tree.

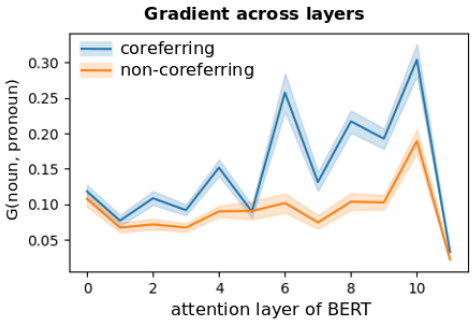

Figure 1: Average gradient $G(noun,pronoun)$ of coreferring vs. non-coreferring pronoun wrt. noun.

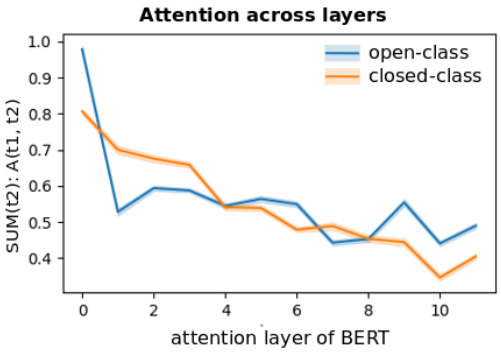

Figure 2: Average attention weights $A(t_1, t_2)$ summed over all t_2, with t_1 an open-class vs. closed-class token.

4 Results

Coreference has a highly significant effect on information flow for each of the three measures defined. I test this with an unequal variance t-test, simply comparing the means of the relevant measure between coreferring and non-coreferring items (all results reported here are highly significant). The attention-based measure shows the greatest effect of coreference (t-statistic: 23), followed by gradient G per layer (t: 19), then gradient G^* with respect to input (t: 13). That is, nouns influence coreferring pronouns more than non-coreferring pronouns. Interestingly, as shown in Figure 1 for G, there is considerable variation across layers (highly consistently across sentences), with peaks for coreference in layers 6 and 10 (0-based indexing). (By contrast, attention A, not shown here, has main peaks at layers 4 and 8.)

Regarding **parts of speech**, open-class items generally provide more information than closed-class items, i.e., they receive more attention (t-statistic: 5), and gradients with respect to them are greater (t: 10). This effect is greatest in the first attention layer (layer 0). With the exception of this layer the gradients for open-class and closed-class words go neatly hand in hand. The attention measure shows a bit more variation in this regard (Figure 2): with the exception of layer 0, the importance of closed-class words seems to decline more steeply than the importance of open-class words.

Lastly, concerning **dependency trees**, there's a correlation between token distances in the gold dependency tree and each of the three measures, particularly gradients, but the correlation is a little under that of a simple right-branching baseline. (Note that the aim here is not to have a dependency parser; e.g., Hewitt and Manning (2019); note in particular that I discard the token representations, and work only with the (much lower-dimensional) attention weights and gradients.) This is not entirely unexpected: BERT is sensitive to sequential order, and words tend to be influenced by nearby words (as can be seen for instance in a heatmap for either of the three measures). Pearson correlations on subsets of tokens, e.g., verb plus core arguments, did not reveal any pattern, but a more fine-grained analysis is ongoing.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154). This paper reflects the author's view only, and the EU is not responsible for any use that may be made of the information it contains.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0. LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51:581–612.