WHAT DOES BERT LEARN ABOUT PROSODY?

Sofoklis Kakouros[a] and Johannah O'Mahony[b]

[a]University of Helsinki, [b]University of Edinburgh
sofoklis.kakouros@helsinki.fi, johannah.o'mahony@ed.ac.uk

ABSTRACT

Language models have become nearly ubiquitous in natural language processing applications achieving state-of-the-art results in many tasks including prosody. As the model design does not define predetermined linguistic targets during training but rather aims at learning generalized representations of the language, analyzing and interpreting the representations that models implicitly capture is important in bridging the gap between interpretability and model performance. Several studies have explored the linguistic information that models capture providing some insights on their representational capacity. However, the current studies have not explored whether prosody is part of the structural information of the language that models learn. In this work, we perform a series of experiments on BERT probing the representations captured at different layers. Our results show that information about prosodic prominence spans across many layers but is mostly focused in middle layers suggesting that BERT relies mostly on syntactic and semantic information.

Keywords: language model, BERT, prosody, prominence, part-of-speech

1. INTRODUCTION

Pre-trained language models (LMs) have achieved high performance on several natural language processing (NLP) tasks such as constituency parsing [1], semantic role labeling, and coreference resolution [2]. Recently, prosody and prosodic phenomena have witnessed increasing attention in natural language applications enabled by pre-trained LMs leading to improved performance [3, 4]. Despite the gains, we do not yet fully understand what enables these models to perform at this level. Their performance suggests that their learning objectives potentially teach the models details about the structure of the language. It remains, however, unclear, what specific prosodic information these models are able to implicitly capture during their pre-training. In this work, we focus on analyzing the prosodic information in one of the most widespread models, BERT (Bidirectional Encoder

Representations from Transformers) [5].

Several recent works have investigated the representations learned at different layers of BERT in an attempt to interpret and understand the linguistic information captured by the model. These studies have uncovered that BERT indeed learns various aspects of the language. For instance, [6] showed that BERT layers capture a rich hierarchy of linguistic information spanning from surface features in the lower layers to syntactic in middle layers and semantic in the higher layers. These findings are further supported by a number of other works with the general observation indicating that learned representations vary with increasing network depth, with greater depth typically involving linguistic functions that require larger contextual relationships across the word tokens [2, 7, 8]. However, to the best of our knowledge, BERT has not been examined with respect to its prosodic information.

In general, prosody can be viewed as the characteristics in an utterance that extend individual phonetic segments and encapsulate phonetic and phonological properties that are not due to the choice of individual lexical items, but depend on factors such as their semantic and syntactic relations [9, 10]. These characteristics convey information about the meaning and structure of an utterance. Although prosody is a characteristic of spoken language, the prosodic patterns in speech are connected and interact with their associated sequences of syllables, words, and phrases. Therefore, it is meaningful to assume that some aspects of the prosodic variation can be captured from text alone [11, 3].

In this work, we investigate how prosodic information is linguistically encoded by probing BERT with respect to the prosodic phenomenon of prominence. Prosodic prominence is defined as the subjective impression of a linguistic unit standing out of its context [12, 13, 14]. Given the recent success of BERT in predicting prosodic prominence [3], we attempt to answer the question of what BERT learns about prosody during its pre-training. Does the model rely on general linguistic and

syntactic knowledge for its prosodic predictions or does it have a different pattern of weight allocations across its layers suggesting that the model can capture prosodic information?

We use three datasets with prominence annotations and examine the weights at different layers and compare them with existing findings from the literature on other tasks. To further validate our approach with the literature we also extract part-of-speech tags for our data and examine the learned layer weights. The code to reproduce the results is publicly available at github.com/skakouros/bert-prosody. In the next we describe the BERT model architecture, related work, experimental methodology and results.

2. BERT

BERT [5] is a language model based on the Transformer architecture [15] that enables the bidirectional pre-training of representations by jointly conditioning on the left and right context in all layers. This allows the model to learn the entire surrounding context of a word which also means that the same word in different context will have a distinct representation. In contrast, earlier approaches looked at text sequences from left to right or by combining left to right and right to left training. The BERT representations are optimized based on two training objectives: (i) predicting randomly masked words in the input, and (ii) predicting whether the next sentence is the subsequent sentence in the input.

Our experiments are based on the bert-base-uncased variant of BERT. The model consists of 12 layers, each with an embedding size of 768, and 12 attention heads.

3. RELATED WORK

Probing network layers to investigate the structural knowledge of the language that a model has captured is an active research area that falls under the topic of neural network interpretability. In recent years there has been an increasing number of studies examining the representations that language models learn. Some works use probing tasks to unveil the linguistic features encoded in neural models [6, 16], others use attribution methods such as Integrated Gradients [17, 18] and some analyze Transformers' attention heads for evidence of linguistic and syntactic phenomena [19]. In this work, we identify each layer's contribution to a specific prosodic task by attaching one trainable weight on each layer of BERT and training a light-weight classification head on top of a frozen pre-trained BERT. This enables us
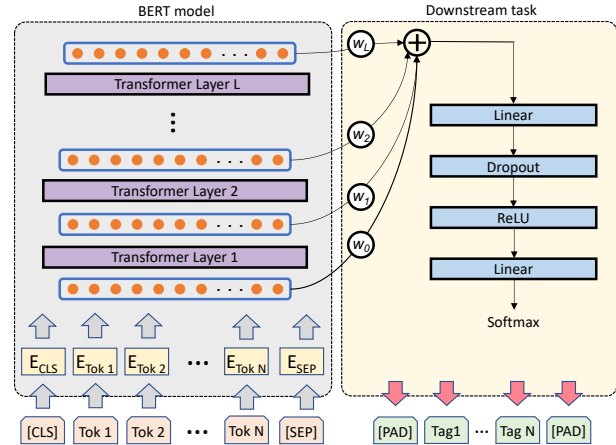

Figure 1: Overview of the experimental setup.

to observe the weight allocations across the layers and compare them with existing findings.

4. LAYER-WISE ANALYSIS

4.1. Layer weights

We obtain the contribution of BERT layers to the prediction task by introducing learnable scalar weights attached to each transformer layer of the model. An overview of the experimental setup can be seen in Fig. 1. We take representations from all transformer layers in the model and collapse them to one via a weighted average. There is one weight for each layer (a total of $L+1$) and all weights are trained jointly with the classification network. The weights are implemented with a learnable vector of size $L+1$, followed by the Softmax function.

4.2. Layer embeddings

To further probe the contribution of individual BERT layers, we extract the embeddings from each layer separately and use them to train a classification head consisting of a single dense layer followed by the Softmax function. For each embedding layer, we then obtain the classification accuracy for the task.

5. EXPERIMENTS

In our experiments we use three datasets: two consisting of read speech and one of spontaneous dialogue speech. These are presented next followed by a description of the experimental setup.

5.1. Data

5.1.1. BURNC

The Boston University Radio News Corpus (BURNC) is a corpus of professionally read news data in American English [20]. The corpus consists of speech from seven speakers (three female). The corpus also contains phonetic alignments, orthographic transcriptions, part-of-speech tags, and prosodic labels. In this work we use the text prompts
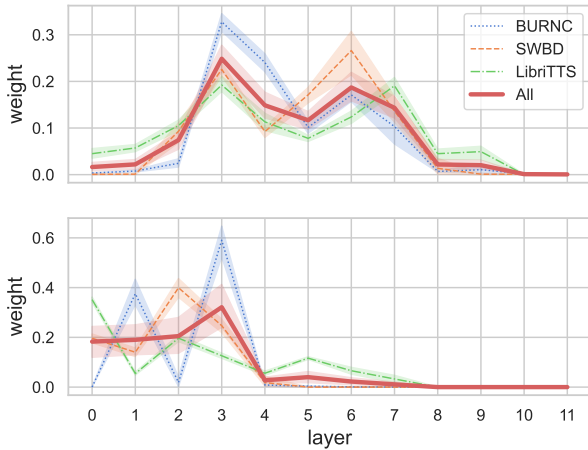
Figure 2: BERT layer weights for prominence (top) and POS (bottom) prediction.

and prosodic labels from the manually labelled part of the corpus (six speakers; approximately 3 h of data). The prosodic labeling system in BURNC is based on the Tones and Breaks Indices (ToBI) labeling convention and includes prosodic phrasing, phrasal prominence, and boundary tones. To obtain a single prominence label, all ToBI pitch accent types (e.g., H*,L*,L*+H) were marked as prominent while the rest as non-prominent.

5.1.2. NXT Switchboard

The NXT Switchboard corpus is a dataset that includes the original Switchboard corpus annotations [21] into one coherent integrated format (NITE XML; NXT) enriched with annotations of prosody and contrast as well as syllable and phone information [22]. The prosody annotations are available for a subset of the data that includes 76 conversations labelled with the ToBI transcribing convention. Similar to BURNC dataset, we use a binary prominence distinction marking words with ToBI accents as prominent and the rest as non-prominent.

5.1.3. LibriTTS

The LibriTTS corpus [23] is a processed (automatically aligned, segmented, and filtered) subset of the original audio and text data of the LibriSpeech corpus [24] that is based on English audiobooks of the LibriVox project. From the corpus we use the *clean* subset that consists of 262.5 hours of read speech from 1230 speakers that were subsequently automatically labelled for prominence in the Helsinki Prosody Corpus (HPC) [3]. From the HPC data, we use the binary prominence tags.

5.2. Experimental Setup

To obtain the layer weights, we train the network (only the classification head and layer weights) with

frozen pre-trained BERT for prominence and POS prediction for each dataset separately. For the training we use a 80-15-5 split for train, validation, and test. We run the training for 20 epochs with a batch size of 4 and we repeat each experiment five times. The results are averaged across the independent runs for each task. The network is configured with a learning rate of $5e-5$ and a different learning rate $1e-2$ for the layers weights. This was done in order to allow the model to adjust the weights for the different layers more rapidly. We extract the weights from the model checkpoint with the best development accuracy and average them over the five runs. We repeat the same procedure and setup for POS prediction.

In addition to the frozen pre-trained BERT we also fine-tune the entire model and report the results in Table 2 to compare the overall model performance with both fine-tuned and frozen BERT models. We use the same learning rates and epochs as the previous experiment with frozen BERT.

For POS prediction we used Spacy [25] to extract the part of speech categories. This resulted into 17 discrete classes. POS classes include, for example, adjectives, adpositions, adverbs (see [25] for a complete list of the coarse POS categories).

Finally, to balance the data and enable better comparison, each dataset is post-processed to include one full sentence per sample. As BURNC may include an entire paragraph and Switchboard several sentences per dialogue turn within a sample, we explicitly set sample size to be one sentence. Thus, a batch size of four includes four sentences.

6. RESULTS AND DISCUSSION

We report overall model performance for POS and prominence prediction in Table 2, layer-specific performance in Table 1, and illustrate how layer weights vary with respect to different tasks in Fig. 2. We are interested in examining how different layers contribute to prominence and POS prediction with respect to findings on other tasks that have indicated different linguistic functions associated with different layers. Overall, for prominence we observe widespread distribution of the layer weights while POS appears more focused in the earlier BERT layers. These are presented in more detail in the next sections.

6.1. Prosodic Prominence

For prosodic prominence, the three datasets tested have shown differences in their overall performance. For example, for frozen BERT, prominence prediction accuracy was 87.14% for BURNC, 78.10% for Switchboard, and 82.67% for

Table 1: Layer-wise accuracy for the test set runs for prominence and POS prediction with frozen BERT. Numbers in bold denote the two top results in each task.

Layer	BURNC		SWBD		LibriTTS	
	Prom	POS	Prom	POS	Prom	POS
0	81.51	93.84	75.32	**94.62**	80.18	89.50
1	83.52	**93.90**	**76.14**	**94.67**	**80.32**	**90.08**
2	85.20	**94.33**	**76.01**	94.02	80.29	**89.53**
3	85.20	93.84	75.47	93.50	**81.08**	89.00
4	**85.26**	93.47	75.25	92.52	80.28	88.40
5	**85.26**	93.17	75.82	91.87	80.00	87.29
6	84.13	92.98	75.71	90.95	79.71	86.12
7	84.46	91.76	75.82	90.11	79.40	84.86
8	83.79	91.03	75.45	88.37	79.05	83.48
9	82.92	89.20	74.71	87.20	78.63	82.36
10	82.59	88.47	74.29	86.64	78.32	81.14
11	83.26	87.19	73.08	82.09	77.27	76.67

LibriTTS. These differences are likely due to the different speaking styles involved in the data, with dialogue speech having the lowest performance. Another interesting observation in the results is that the model performance is degraded for the fine-tuned models. It seems that when fine-tuning the model, generalizability decreases due to overfitting on the idiosyncrasies of the training data. With the frozen model, performance is high and on a par with results reported in the literature. For instance, for LibriTTS [3] report 83.20% and in our setup we obtained 82.67%.

When it comes to layer weights, prosodic prominence has a widespread pattern of weights across BERT layers. Weights span most layers and are primarily focused within layers 2-8 with a peak appearing at layer 3. Interestingly, the pattern of weights appears to be quite similar across the three datasets tested. The spread of the weights suggests that different types of linguistic information are used in the prediction of prominent tokens. One interpretation of the results is that the model relies greatly on surface linguistic features such as POS but also involves syntactic and semantic information with increasing layers (see also [6]). For BURNC, we also observe that the third layer has a high weight for both prominence and POS prediction. It is possible that prominence in BURNC (professionaly read speech) correlated more with POS than audiobooks (LibriTTS) and dialogue speech (SWBD). Another possibility for these differences could be also attributed to the different prominence coding schemes in the three datasets.

6.2. POS

Part-of-speech information seems to be encoded in the early BERT layers with model accuracy being high for both frozen and fine-tuned runs of the experiments and across all datasets tested. Fine-tuning the model leads to improved performance,

Table 2: Accuracy for the test set runs for prominence and POS prediction with frozen and fine-tuned (ft) BERT.

Prominence	BURNC	SWBD	LibriTTS
frozen	87.14	78.10	82.67
ft	85.53	75.95	80.32
POS			
frozen	95.97	97.94	98.49
ft	97.56	98.54	98.96

where for BURNC we get an increase in accuracy from 95.97% with the frozen model to 97.56% with fine-tuning. Switcboard and LibriTTS perform similarly with an increase in performance when the model is fine-tuned.

Layer weights for POS demonstrate a very different pattern when compared to prominence. POS information is found mainly in the lower layers of BERT with weights across the datasets varying but being focused in the early layers of the model. Most of the information seems to come from layers 0-4 which have been shown to encode surface features [6]. This finding is also in agreement with other work that shows maximum POS tagging performance in the lower layers of models with accuracies ranging from 97.2% to 97.4% [7, 2].

7. CONCLUSIONS

In this work, we performed a series of experiments on prosodic prominence to investigate whether prosody is part of the structural information of the language that BERT learns. Our results show that BERT captures information about prosodic prominence through a widespread allocation of weights across its layers reaching high performance. The weight allocations suggest that BERT relies on a variety of linguistic information for its predictions including surface linguistic features such as POS but also involving syntactic and semantic information. In future work, we will explore the same tasks with an extended set of datasets and methodological approaches. In addition to layer weights, we want to include layer integrated gradients as an attribution method in the experiments. We also aim to examine differences between the styles in the datasets, that is, read versus dialogue speech.

8. ACKNOWLEDGEMENTS

9. REFERENCES

[1] N. Kitaev and D. Klein, "Constituency parsing with a self-attentive encoder," *arXiv preprint arXiv:1805.01052*, 2018.

[2] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[3] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," in *Proceedings of NoDaLiDa*, 2019.

[4] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "Bert, can he predict contrastive focus? predicting and controlling prominence in neural tts using a language model," *arXiv preprint arXiv:2207.01718*, 2022.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[7] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih, "Dissecting contextual word embeddings: Architecture and representation," *arXiv preprint arXiv:1808.08949*, 2018.

[8] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," *arXiv preprint arXiv:1905.05950*, 2019.

[9] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes*, vol. 25, no. 7-9, pp. 905–945, 2010.

[10] S. Kakouros and O. Räsänen, "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," *Cognitive science*, vol. 40, no. 7, pp. 1739–1774, 2016.

[11] S. Kakouros, J. Pelemans, L. Verwimp, P. Wambacq, and O. Räsänen, "Analyzing the contribution of top-down lexical and bottom-up acoustic cues in the detection of sentence prominence," *Proceedings Interspeech 2016*, vol. 8, pp. 1074–1078, 2016.

[12] S. Kakouros and O. Räsänen, "3pro–an unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[13] J. Terken and D. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and experiment*. Springer, 2000, pp. 89–127.

[14] S. Kakouros and O. Räsänen, "Automatic detection of sentence prominence in speech using predictability of word-level acoustic features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, S. R. Bowman, D. Das *et al.*, "What do you learn from context? probing for sentence structure in contextualized word representations," *arXiv preprint arXiv:1905.06316*, 2019.

[17] S. Ramnath, P. Nema, D. Sahni, and M. M. Khapra, "Towards interpreting bert for reading comprehension based qa," *arXiv preprint arXiv:2010.08983*, 2020.

[18] A. Nayak and H. P. Timmapathini, "Using integrated gradients and constituency parse trees to explain linguistic acceptability learnt by bert," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, 2021, pp. 80–85.

[19] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.

[20] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[21] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[22] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, no. 4, pp. 387–419, 2010.

[23] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[25] M. Honnibal and I. Montani, "spaCy 3: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.