Where exactly does contextualization in a PLM happen?

Soniya Vijayakumar, Tanja Bäumel, Simon Ostermann and **Josef van Genabith**
German Research Instutite for Artificial Intelligence (DFKI), Saarland Informatics Campus, Germany
soniya.vijayakumar, tanja.baeumel, simon.ostermann, josef.van_genabith @dfki.de

1 Introduction

Pre-trained Language Models (PLMs) have shown to be consistently successful in a plethora of NLP tasks due to their ability to learn contextualized representations of words (Ethayarajh, 2019). BERT (Devlin et al., 2018), ELMo (Peters et al., 2018) and other PLMs encode word meaning via textual context, as opposed to static word embeddings, which encode all meanings of a word in a single vector representation. Much recent work has explored understanding the syntactic, semantic or contextualization information present in the hidden-states or attention-heads of the various PLMs by extracting contextualized representations either from the last output layer or concatenating the mean of representations across all layers (Khattab and Zaharia, 2020; Ravfogel et al., 2020; Zhao et al., 2020; Khattab and Zaharia, 2020; Wiedemann et al., 2019; Ethayarajh, 2019; Mareček and Rosa, 2019; Clark et al., 2019). In this work, we present a study that aims to localize where exactly in a PLM word contextualization happens. In order to find the location of this word meaning *transformation*, we investigate representations of polysemous words in the basic $BERT_{uncased}$ 12 layer architecture (Devlin et al., 2018), a masked language model trained on an additional sentence adjacency objective, using qualitative and quantitative measures.

Our main contribution is that unlike previous work, we do not restrict ourselves to the output layer(s) of such networks, but also investigate different *sub-layers* within each BERT encoder layer. We also apply Dimensionality Reduction (DR) techniques on these latent sub-layers to better visualize our qualitative findings.

First, we confirm findings of Ethayarajh (2019) and other works, indicating that higher layers of BERT exhibit higher degrees of contextualization. Second, by investigating contextualization also in *encoder sub-layers*, we find for the first time that representations in the self-attention sub-layers exhibit stronger and earlier signs of contextualization as compared tothe activation and output sub-layers.

2 Methodology

Experimental Setup: We use the Contextualised Polysemy Word Sense v2 Dataset (Haber and Poesio, 2020) which contains custom samples of polysemous words in sentential contexts. We feed two sentences for each polysemous word to BERT and extract sub-layer vector representations for each encoder layer. For each word, we arrive at a set of vectors: *Self-Attention (SA) sub-layer, Activation (Acts) sub-layer, and Output sub-layer*. We also extract the BERT static word embeddings (layer-0) for understanding semantic divergence of the sub-layers.

Measures of Contextuality: *Sub-Layer Similarity.* Let *w* be the polysemous word that appears in a pair of sentences $\{s1, s2\}$ at index *i,j* in its respective sentences, $\{x_{s1}, x_{s2}\}$ the sub-layer vector representations of the model *m*. The sub-layer similarity of word *w* in layer *l* is:

$$SubLayerSim_x(w^l_{i,j}) = \frac{cos(x^l_{i_{s1}}, x^l_{j_{s2}})}{||x^l_{i_{s1}}||||x^l_{j_{s2}}||} \quad (1)$$

where *x* is Self-Attention sub-layer, Activation sub-layer and Output sub-layer or $l = \{0, 1, ..., 11\}$.

Static Word Embedding Similarity. For each word, we determine the cosine similarity between each sub-layer and its respective static word embedding from layer 0 (denoted as *WESim*).

Principal Components Analysis (PCA). For better visualization, we reduce the High-Dimensional (HD) sub-layers (12 x 768, 12 x 3072) into two principal components using the PCA technique. PCA preserves the actual relative distance in the Euclidean data space and Principal Components (PCs) capture the direction of maximum variance.

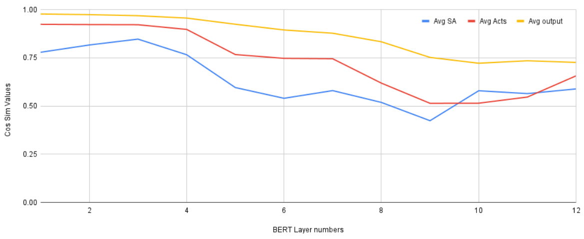

Figure 1: Layer-wise average Self-Attention (SA), Activations (Acts) and Output *SubLayerSim* for all words.

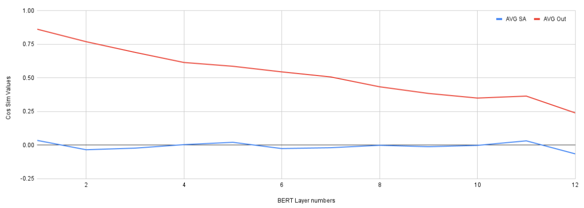

Figure 2: Layer-wise average Self-Attention (SA) and Outputs *WESim* for all words. the Acts sub-layer similarity is not reported due to the dimensionality mismatch.

We determine squared L2 distances of the PCs between reduced sub-layer vector representations in the pair of sentences. We use these distance measures to quantitatively conform the observations made using *SubLayerSim* similarities.

Intuitively, the higher the *SubLayerSim* between the polysemous words, the closer the respective vector representations are to each other, hence they are less contextualized. The lower the *WESim*, the further the vector is away from the (uncontextualized) static word embedding. On the other hand, the higher the *L2 Distance*, the further the respective words (PCs) are from each other. This implies the actual relative distance in the higher dimension is high, indicating more contextualization.

3 Results & Discussion

Examining the average *SubLayerSim* of all polysemous words, we observe that the Output sub-layer similarity is closer to one, indicating a lower degree of contextualization as compared to Acts and SA sub-layers (Figure 1). The Acts and SA sub-layers *SubLayerSim* are closer to each other, with higher contextualization in the SA sub-layer. The all-words average *SubLayerSim* for the sub-layers indicate similar observations (Table 1).

	Avg Sa	Avg Acts	Avg Outs
SLSim	0.6329	0.7309	0.8614
WESim	0.008	-	0.521
L2 Dists	3.217	3.413	1.195

Table 1: SLSim: SubLayerSim, WESim Word Embeddings and L2 distances for each sub-layer and all words.

BERT's layer-wise average *SubLayerSim* of all sub-layers decreases towards higher layers (lowest at layer 9), indicating the existence of higher contextualization layers between layers 5 and 9 (Figure 1). This observation is similar to Ethayarajh (2019)'s experiments: The upper hidden layers of contextualizing models produce more context-specific representations. An interesting observation is that the sub-layer similarities become higher and closer to each other in the penultimate layer of BERT (layer 12), which could be an indication of the model's sensitivity to output form and training objective. A notable finding in WESim is that, SA sub-layer WESim remains relatively consistent across BERT Layers where as the Output sub-layer WESim consistently decreases (Figure 2). This indicates the influence of residual connections in the BERT encoder Output sub-layer.

PCA: We observe that the PCA bi-plots for SA, Acts and Output sub-layers are structurally different, indicating different structural alignment in their respective HD spaces. Examining the per-word average L2 distances, we observe that the Output sub-layer L2 distances are much lower than the respective SA and Acts sub-layers, indicating a stronger contextualization in the SA and Acts sub-layers (Table 1).

4 Conclusion & Future (on-going) research

In our research, we dive deeper into *sub-layers* of each BERT encoder layer to localize the contextualization of polysemous words. Using contextuality measures and PCA, we observe that the contexts of these words are strongly captured in the Self-Attention sub-layer compared to the respective Activation and Output sub-layers and in higher layers of BERT. In other words, we find word meaning *transformation* to occur strongly in SA and Acts sub-layer as compared to the Output sub-layers. We are currently working on exploring these methods on benchmark WSD corpora (Raganato et al., 2017) on other contextualized models. Our future research involves applying Representational Similarity Analysis (RSA) across the sub-layers (Abnar et al., 2019) to understand if contextualization redundancies exist in these sub-layers.

Limitations

In this work, we are faced with a few limitations. First, the available WSD benchmark datasets are not in the format we require and hence, we conducted our experiments with a relatively smaller dataset, as presented above. Second, with the limited dataset, all the sentences have a standard structure for each polysemous word, i.e, each polysemous word is in the second position after the stopword 'The'. We are yet to experiment with sentences that have different indices for polysemous words. Third, we use two different metrics, cosine similarity and euclidean distance on a PCA for understanding the extent of contextualization. These two metrics have different properties making them not directly comparable. *SubLayerSim* measure helped us in quantifying the contextualization whereas we used PCA as a visualization technique, to qualitatively confirm our observations.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. *CoRR*, abs/1909.00512.

Janosch Haber and Massimo Poesio. 2020. Word sense distance in human similarity judgements and contextualised word embeddings. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 128–145, Gothenburg. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

David Mareček and Rudolf Rosa. 2019. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Jacob Goldberger, and Yoav Goldberg. 2020. Unsupervised distillation of syntactic information from contextualized word representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 91–106, Online. Association for Computational Linguistics.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings.

Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. Quantifying the contextualization of word representations with semantic class probing.