The Universe of Utterances According to BERT

Dmitry Nikolaev **Sebastian Padó**
Institute for Natural Language Processing, University of Stuttgart
dnikolaev@fastmail.com pado@ims.uni-stuttgart.de

Abstract

It has been argued that BERT "rediscovers the traditional NLP pipeline", with lower layers extracting morphosyntactic features and higher layers creating holistic sentence-level representations. In this paper, we critically examine this assumption through a principle-component-guided analysis, extracing sets of inputs that correspond to specific activation patterns in BERT sentence representations. We find that even in higher layers, the model mostly picks up on a variegated bunch of low-level features, many related to sentence complexity, that presumably arise from its specific pre-training objectives.

1 Introduction

The Transformer architecture of neural networks (Vaswani et al., 2017) shows state-of-the-art performance on a range of NLP tasks (Wang et al., 2018, 2019). At the same time, the question of what Transformer models learn exactly has motivated a number of studies into the representations that they construct (Rogers et al., 2020; Chi et al., 2020; Papadimitriou et al., 2021), with an increasingly popular answer being that they recreate the classical NLP pipeline of incremental abstraction, from morphosyntax to semantics (Tenney et al., 2019; Geva et al., 2021).

In this paper, we put this finding to the test, asking to what extent representations learned by pre-trained BERT (Devlin et al., 2019) capture systematic meaning distinctions, as opposed to more shallow and potentially idiosyncratic properties. The challenge of this question is that it is open-ended: in order not to bias the analysis, we do not want to rely on a set of categories that we *a priori* expect to be relevant, in contrast to most probing approaches, which correlate model representations with properties of inputs or performance metrics (see Section 2 for details).

Instead, we adapt the approach that Geva et al. (2021) proposed to analyze decoder-only Transformer models with causal masking. They regard feed-forward (FF) sublayers in such models as *neural memory units* and extract inputs that produce maximal activations in a random subset of their neurons. Manual analysis of these sets shows what categorization of inputs arises inside the model.[1]

We extend the approach by Geva et al. in two dimensions. First, we apply it to pre-trained bidirectional encoder-only transformer models like BERT (Devlin et al., 2019; Liu et al., 2019), which, unlike causal LMs, do not have a specific token guaranteed to represent all of the input. To do so, we analyze two types of "prominent" tokens: the CLS pseudo-token, often used for whole-sentence representation (Ma et al. 2019; even if its usefulness for downstream tasks is debatable, cf. Reimers and Gurevych 2019), and the first subword of the `root` element in sentences annotated with Universal Dependencies (Nivre et al., 2020). We regard these two tokens as good candidates for loci of high-level, abstract representations of inputs learned by BERT.

Second, we replace the analysis of random neurons by *guided exploration*. We find that embeddings of both CLS tokens and `root` tokens at upper layers are highly intercorrelated. Therefore we propose to analyze major principal components of activation matrices, in essence tracking influential groups of highly congruent neurons.

We exploit this approach to provide an analysis of the sentence patterns that BERT attends to. We find that while lower levels of BERT are predictably more attuned to lexical effects, activations in higher levels track a wide range of idiosyncratic phenomena from various linguistic levels, from individual wordforms and bigrams to lexical classes (rare

[1] An automated approach to finding features that trigger neuronal activations was proposed by Rethmeier et al. (2020). They assign probability distribution over features to different neurons, which makes qualitative analysis impractical.

Proceedings of the 15th International Conference on Computational Semantics pages 99–105
June 21–23, 2023.

words), syntactic patterns (e.g., clauses with imperatives), and miscellaneous sentence types (recommendations, incomplete sentences, short exclamations). Overall, our results indicate that the typology of sentences according to BERT is dominated by what may be called *natural classes* (Mielke et al., 2011) – clusters of objects that are characterized by a combination of values of several features – with the embeddings showing little evidence of principled semantic properties.[2]

2 Related Work

Probing transformers Two prominent avenues of the study of Transformers in NLP are (i) probing analysis of internal representations of linguistic inputs computed by the models (e.g., Vulić et al., 2020; Pimentel et al., 2020; Belinkov, 2022) and (ii) the analysis of the attention patterns that Transformers converge on to compute these representations (Voita et al. 2019; Bian et al. 2021 and many others). Both strategies rely on predefined arrays of NLP tasks and features, which are either used as benchmarks or are selected to highlight peculiarities of models.

In contrast, Geva et al. analyze the representation of the last unmasked token of the input sequence in the causal language model by Baevski and Auli (2019), which serves as the representation of the whole prefix. By sampling neurons from feedforward sublayers of the model and manually inspecting sentences that give rise to maximum values of these neurons, they show that the latter recognise different patterns in the input – with lower-layer activations tuned to more superficial lexical and syntatctic features and upper layers arguably more tuned to semantics. We extend this approach to bidirectional LMs.

3 Methods

3.1 Models and Data

All experiments are conducted based on the `bert-base-cased` model provided by Wolf et al. (2020). We use the train and development splits of the Georgetown University Multilayer (GUM) corpus (Zeldes, 2017), which is annotated with Universal Dependencies. Together, the two splits comprise 6,507 sentences.

[2]The code used for the analyses in this paper is available at `https://github.com/macleginn/universe-of-utterances`

3.2 Analysis Procedure

Token selection and representation We consider two tokens that are promising candidates as loci for high-level categories that we would expect BERT pre-training to extract from inputs: the CLS token and the dependency root of the sentence.

For the CLS token, used in pre-training for the next-sentence-prediction (NSP) task, we concentrate on the output of the pooler layer: an additional MLP is applied to the raw BERT encoding before it is fed into the classifier head.

The `root` token does not play a special role in pre-training, but we assume that, as it largely corresponds to the head predicate of the clause, it should attend to its various syntactic elements in order to be selected correctly and to guide selection of other tokens.[3] To analyze `root` tokens, we experiment with the outputs of feedforward sublayers in the 3rd, 6th, and 11th BERT layers, which should roughly correspond to different layers of generic linguistic abstraction attained by the model. The final layer has been suspected of being too task-specific (Kovaleva et al., 2019).

Analysis procedure Our analysis proceeds in two steps for both types of tokens: (1) we gauge the extent of redundancy in the representations, given that BERT neurons are known to be highly redundant in general (Dalvi et al., 2021). As we will show in Section 4.1, CLS-token embeddings are in particular highly redundant. Consequently, (2) we identify the first 5 principal components of embedding matrices and extract sentences with maximum and minimum scores for these PCs to manually to identify shared features, similarly to Geva et al. (2021).

Hypotheses Under the theory that BERT rediscovers the traditional hierarchy of NLP tasks during pre-training (Tenney et al., 2019), we expect it to be possible to interpret principal components as bundles of linguistic features, with higher layers moving from morphosyntax towards sentence semantics. Alternatively, we can hypothesize that BERT optimizes its representations primarily for its pre-training objectives (next-sentence prediction for CLS and masked-token prediction for `root` tokens), which would presumably not support a clean interpretation in terms of a feature hierarchy.

[3]Another vector that is often used as a stand-in for the whole sentence is the average of all tokens embeddings. As by construction it cannot be tied to any particular sentence component, it is less interpretable.

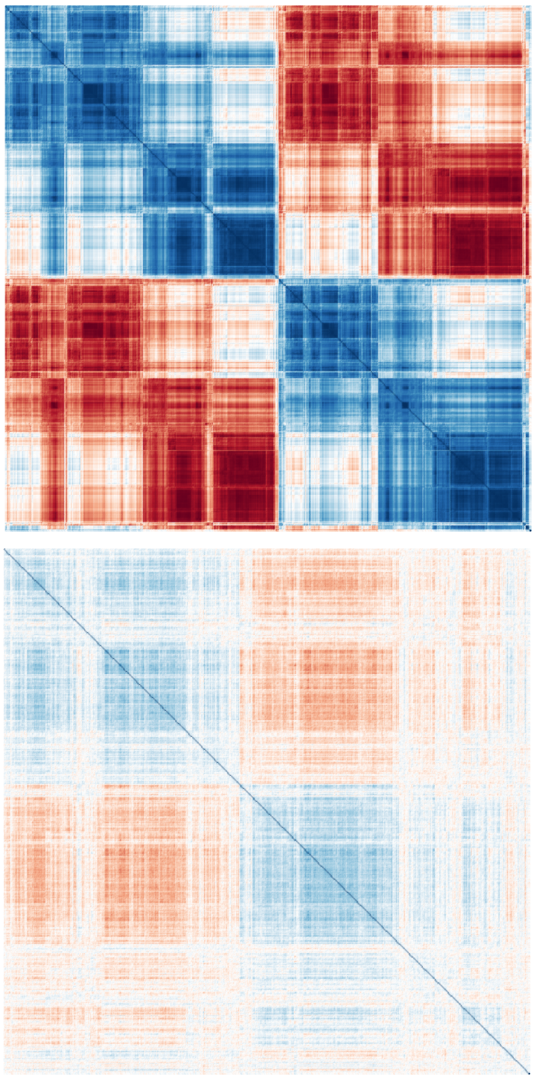

Figure 1: Correlations of neuronal activations in the output of the pooler layer for CLS tokens (top) and the output of the FF sublayer in layer 11 for `root` tokens (bottom). Rows and columns were reordered using hierarchical clustering. More intense blue/red hues correspond to stronger positive/negative correlations.

4 Results

4.1 Neural Redundancy and Major Principal Components

As motivated above, we first assess the redundancy of the pre-trained BERT embeddings for `root` tokens and CLS tokens (Dalvi et al., 2021). The results, shown in the correlation plots in Figure 1, reproduce the findings of earlier studies: there are evident clusters of mildly correlated and anticorrelated neurons in `root` token embeddings, while CLS neurons are extremely intercorrelated.

This redundancy motivates our use of principal-components analysis (PCA) to reduce the dimensionality of these representations. We find that more than 50% of the variance in the output of the pooler layer for CLS is explained by the first component alone; the first three components explain 76% of the variance. In the case of FF sublayer embeddings of `root` tokens, the first component explains 25% of the variance in the 11th layer (36% for the first 3 PCs together), but only 5% in the 6th layer (10% for the first 3 PCs), and only 3% in the 3rd layer (8% for the first 3 PCs). Taken together, this shows that as BERT progresses in its analysis of the inputs, it aggressively discards more and more information (Tishby and Zaslavsky, 2015). Upper layers are less redundant than middle layers (Dalvi et al., 2020) but might still be overparameterized (although this may also be interpreted as "spare capacity" for fine-tuning).

4.2 Principal Component-based Analysis

Sentences with maximum and minimum scores for 5 PCs for all studied settings can be found in the paper's code repository.[4]

4.2.1 CLS Tokens

Table 1 shows examples and statistics for the first 5 PCs of the CLS embedding space. The PCs may be interpreted as as largely corresponding to *sentence complexity*: they are noticeably correlated with sentence length and somewhat correlated with the number of rare words, operationalized as hapax legomena in the test set. The particular patterns, however, are highly varied.

Sentences with top scores on **PC 1** are all short, consist of bare NPs, and do not include rare words. Sentences with minimal scores are, by contrast, mostly long and include rare words, such as person and place names.

Examples with minimal values for **PC 2** are all short conversational utterances, while sentences with maximum values do not form a coherent group.

Values for **PC 3** demonstrates the highest correlation with sentences length (0.57). Examples with minimum values are all short quotes without verbs of (reported) speech. Examples with maximum values seem all to be narrative sentences with first-person-pronoun subjects.

Minimum values for **PC 4** are mostly triggered by sentences with an opening quote mark but without a closing one, i.e. those starting a direct-speech

[4] https://github.com/macleginn/universe-of-utterances

PC	SL	HL	Inputs w/ extremal values
1st	-0.26	-0.1	**[max]** Estimated electricity use in residential sector; Second baseman / Shortstop / Outfielder; High school career
2nd	0.37	0.37	**[min]** Yeah, I bet.; Yeah , that's a good idea.; Probably.; Sure.; Nah, I'm kidding.; Yeah , "think again" or something like that.; I have no idea.
3rd	0.57	0.22	**[min]** "With what?"; "Have you thought of that?"; "Why?"; "Are you ashamed of her?"; "It's not a joke."; "No." **[max]** I would find myself entering those crypts...; ...I came up with an individual story called Thad's World Destruction...; We just want to be able to bring, like she said, bring light into the entertainment...
4th	0.34	0.27	**[min]** We are a colony.; They're going to implant a chip.; Go away!
5th	-0.06	0.1	**[min]** THE END; Chapter Two: Master Lunre; 1 Harvest and prune

Table 1: CLS token analysis: Spearman correlations with measures of complexity (SL: sentence length, HL: hapax-legomena counts per sentence) and examples inputs with extremal values (minimum / maximum).

segment. Sentences with maximum PC 4 values are not easily interpretable.

Minimum values for **PC 5** are shown by a varied set of sentences many of which are chapter/section names. Sentences with maximum values on this axis do not afford a simple interpretation.

Overall, it is evident that CLS representations are finely attuned to different kinds of sentences that are likely to appear in particular contexts and are thus informative for the next-sentence-prediction task. Their semantic properties, which CLS tokens are often assumed to be representations of, seem to be largely irrelevant.

4.2.2 `root` Tokens

Layer 3 As expected, the FF sublayer of layer 3 is focused on shallow features. Sentences with minimum values for **PC 1** are headed by the verb *have*. Minimal values for **PC 2** correspond to a combination of the verb form *said* and quote marks. Maximal values for **PC 3** track non-third-person subject and the verb *know* in the present tense, preferably in combination.[5] **PC 4**, despite being orthogonal to previous components, assigns minimal values to sentences headed with *have* and maximal values to sentences with *said* and quote marks. Similarly, **PC 5** assigns minimal values to sentences with *know* but maximal values to sentences headed by forms of *go* and *come*, including phrasal verbs with widely differing semantics (*go on*, *go through*, *come home*), which shows that this combination is more collocational than semantic.

Layer 6 We expect Layer 6 to encode more abstract features. However, **PC 1** of `root` tokens on layer 6 is highly negatively correlated with sentence length ($r = -0.62$). Examples with high scores include one-word utterances (*Alright.*), dates, and image captions of the form *Image: [AUTHOR].* Minimum values of **PC 2** correspond to sentences with forms of the verb *say* and a couple of other verbs of speech as the head predicate. Maximum values seem to be uninterpretable. Similarly, minimum values of **PC 3** correspond to sentences headed by the verb *have*, while sentences with maximum values are, with several exceptions, headed by *be*. Small values for **PC 4** are indicative of verbs of creation (*make*, *construct*, *build*). Small values of **PC 5** again correspond to sentences with the verb *have*. Sentences with high scores on this component, however, are predominantly headed by a verb in the imperative mood (*see*, *know*, *come*, *tell*, etc.).

Layer 11 We expect Layer 11 to represent high-level semantic features. But again, **PC 1** on layer 11 is also correlated with sentence length, this time positively ($r = 0.57$). This time, sentences with minimal scores have a rather specific form of technical instructions, including recipes.[6] Minimal values of **PC 2** seem to be connected to different kinds of short sentences (*You ass.*; *Absolutely great.*; *She sighs.*), incomplete phrases (*They're really —*; *Melanie lies but —*), and nominative heading-like constructions (*Basalt columns*; *Country-specific advise*). Minimal values of **PC 4** correspond to sentences headed with *there's*, *there is*, *it's*, and,

[5] *Know your audience.*; *We know self-isolation works.*

[6] *Position a large mirror so you can check your positioning and see what you're doing.*; *Add six Skittles to 25 ml of vodka.*

somewhat incongruously, *I'm*. **PCs 3** and **5** do not support an obvious interpretation.

Discussion Overall, the PCs of `root` token representations on layer 3 are oriented towards frequent verbs tokens, while layer 6 adds a morphosyntactic category of imperatives, and layer 11 singles out a wide variety of sentence patterns anchored by features at the level of surface properties (sentence length, presence of a particular verb), lexical groups (verbs of creation), syntactic categories (imperatives), or text types (technical instructions). Sentence length remains a recurring feature, as it is for the CLS token.

5 Conclusions

The good performance of Transformers on downstream tasks is often explained by their ability to extract meaningful linguistic, generalizing features from raw text (Tenney et al., 2019; Rogers et al., 2020; Geva et al., 2021). When approaching this problem from the point of view of a particular set of tasks, however, there is always the danger that good model performance is due to accidental covariates in the data that help models solve the task without creating useful generalizations (Levy et al., 2015; Gururangan et al., 2018).

Our analysis of loci in BERT that are highly likely to aggregate linguistic generalizations about the input sentence indicates that this problem might indeed be present in this model as well: we find a conspicuous absence of high-level generalizations and prominent shallow features even in the final layers, arguably because they prove useful in solving the cloze and next-sentence-prediction pre-training tasks. Many of these are complexity-related, similar to biases found in word embeddings (Wilson and Schakel, 2015).

These findings arguably go some way towards explaining the instability of the performance of different instances of BERT on the same downstream task (McCoy et al., 2020) and of the variance in the effects of BERT interventions (Sellam et al., 2021). The question of whether it is possible to create a pre-training task that would nudge the model towards extracting high-level features remains open.

Limitations

One limitation of this study is that it demands manual inspection of extracted sentences. While this makes it possible to identify patterns in a way not prejudiced by the downstream task or the available annotations of the inputs, it also makes it harder to provide quantitative arguments in favor of the proposed analysis.

Another limitation is that we only focus on maximum and minimum values of the principal components when extracting diagnostic sentences. This provides for a clear interpretation when PCs can be construed as well-defined axes; however, sometimes they appear to be "discontinuous", with different properties surfacing at the two extreme points. This suggests that there may be other interesting classes of inputs encoded by mid-range values.

References

Alexei Baevski and Michael Auli. 2019. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yuchen Bian, Jiaji Huang, Xingyu Cai, Jiahong Yuan, and Kenneth Church. 2021. On attention redundancy: A comprehensive study. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 930–945, Online. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4908–4926, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. 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. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from BERT: An empirical study. *arXiv preprint arXiv:1910.07973*.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Jeff Mielke, Elizabeth C. Zsiga, and Paul Boersma. 2011. The nature of distinctive features and the issue of natural classes. In Abigail C Cohn, Cécile Fougeron, and Mary K M. Huffman, editors, *The Oxford Handbook of Laboratory Phonology*, pages 185–196.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Rethmeier, Vageesh Kumar Saxena, and Isabelle Augenstein. 2020. TX-Ray: Quantifying and explaining model-knowledge transfer in (un-)supervised NLP. volume 124 of *Proceedings of Machine Learning Research*, pages 440–449, Virtual. PMLR.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2021. The MultiBERTs: BERT reproductions for robustness analysis. In *Proceedings of ICLR 2022*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. 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. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Benjamin J. Wilson and Adriaan M. J. Schakel. 2015. Controlled experiments for word embeddings. *CoRR*, abs/1510.02675.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.