Probe-Less Probing of BERT's Layer-Wise Linguistic Knowledge with Masked Word Prediction

Tatsuya Aoyama
Georgetown University
ta571@georgetown.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Abstract

The current study quantitatively (and qualitatively for an illustrative purpose) analyzes BERT's layer-wise masked word prediction on an English corpus, and finds that (1) the layer-wise localization of linguistic knowledge primarily shown in probing studies is replicated in a behavior-based design and (2) that syntactic and semantic information is encoded at different layers for words of different syntactic categories. Hypothesizing that the above results are correlated with the number of likely potential candidates of the masked word prediction, we also investigate how the results differ for tokens within multiword expressions.

1 Introduction

The attention mechanism of Transformers (Vaswani et al., 2017) has enabled language models (LMs) to effectively incorporate contextual information into word representation. One such model, BERT (Devlin et al., 2019), has been shown particularly useful in a wide range of downstream tasks, outperforming the state-of-the-art benchmarks in many cases. However, it is yet to be clear what exactly such LMs learn, and what information is encoded in their contextual word representations (CWRs). For this reason, much work has been devoted to answering these questions, often referred to as BERTology (see Rogers et al., 2020 for a comprehensive review).

Among such studies, of particular interest is the localization of linguistic knowledge. As BERT consists of multiple layers (12 layers for `bert-base` and 24 layers for `bert-large`), it is crucial to understand what information is encoded in each layer, and how it differs from one another. However, the methodologies employed in such studies differ substantially from each other (§2): some directly utilize the internal structure of such models by training probing classifiers, while others study the behaviors of such models at inference time. Structure-based probes have often been successful at assigning particular domains of linguistic knowledge to local regions, yet the reliance on probing classifiers (and the introduction of extra parameters) makes it unclear if such linguistic knowledge is just an artifact of the classifier or is truly encoded in the model. Behavior-based probes do not rely on external classifiers, but tend to focus on qualitative analyses of the outputs from the final layer, whereas quantitative analysis of layer-wise output remains understudied in the behavioral paradigm.

In this study, we explore layer localization with behavioral probing. Specifically, we mask out tokens one at a time and check whether BERT predicts the same word, another word with the same part of speech, or neither (3). By using different layers for the prediction, we can determine which parts of the network correspond to higher or lower rates of congruent predictions. Along with generally confirming some of the main observations of the structure-based probing studies, we find considerable variation by part of speech and some effect of multiword expression status, and discuss possible interpretations of these findings (§4).

2 Previous Work

2.1 Structure-Based

Since the advent of BERT (Devlin et al., 2019), much work has been devoted to revealing what linguistic knowledge it has. Among such studies, Tenney et al. (2019a) observed that one line of work is behavior-based, while the other directly investigates the structure of the CWRs. Whereas the former focuses on the qualitative error analyses of BERT's predictions on certain controlled tasks, the latter directly probes the internal structure of the model. Building on the latter line of the work, Tenney et al. (2019a) apply a probing method called *edge probing* (Tenney et al., 2019b), which allow them to infer what sentence-level information BERT encodes based on a given span by restricting the input to the probing classifier. They find that BERT's layer-wise linguistic knowledge resembles classical NLP pipelines; in other words, lower layers are more responsible for syntactic knowledge and higher layers for semantic knowledge, although syntactic knowledge is more localizable at lower layers whereas semantic knowledge is rather spread across the layers.

Jawahar et al. (2019) make similar observations, based on a suite of probing tasks developed by Conneau et al. (2018). They find that the lowest layer is most successful at phrase detection, and the performance degrades until layer 8, beyond which it reaches a plateau. In another set of experiments, they find that lower, middle, and higher layers are responsible for surface, syntactic, and semantic information, respectively. Corroborating this result, Hewitt and Manning (2019) employ a novel method called a structural probe to retrieve a syntactic tree from contextualized word embeddings, and find that the representation from middle layers have better performance in the tree retrieval task.

With an increasing number of studies employing probing classifiers, in their comprehensive review of BERTology, Rogers et al. (2020) raise a warning that such probing may not provide us with a full picture of what BERT is: "If a more complex probe recovers more information, to what extent are we still relying on the original model?" Indeed, while some studies use a linear classifier as a probe to limit the number of newly introduced parameters (e.g., part of Liu et al., 2019), others use more complex models, such as multi-layer perceptron (MLP), obscuring the source of success on probing tasks. Hewitt and Liang (2019) suggested a metric called *selectivity* to measure how well a probe reflects the actual linguistic knowledge encoded in the CWRs in question, as opposed to learning the task independently of such CWRs.

2.2 Behavior-Based

Complementing such limitation of probing studies, more recent works have attempted to avoid introducing new parameters through creative probing methodologies, such as contextual word embedding (CWE) similarity ranking (Gessler and Schneider, 2021), and direct probe (Zhou and Srikumar, 2021).

In fact, the other line of work, which Tenney et al. (2019a) described as behavior-based, which usually relies on qualitative (error) analyses of BERT's predictions on controlled tasks, is parameter-free and utilizes BERT's behaviors at inference time, and Rogers et al. (2020) also argue for the importance of this line of work. Such work includes investigation of semantic knowledge (Ettinger, 2020; Marvin and Linzen, 2018) and syntactic knowledge (Goldberg, 2019; Poliak et al., 2018).

For example, analyzing BERT's masked word prediction output on controlled tasks developed in psycholinguistic studies, Ettinger (2020) finds that BERT struggles with common sense and pragmatics, role-based event prediction, and negation. Goldberg (2019) also studies BERT's masked word prediction outputs on both naturally occurring sentences and manually crafted stimuli, finding that BERT is sensitive to subject-verb agreement.

While these studies have revealed a great deal about BERT's linguistic knowledge, they have primarily focused on (1) content words, such as verbs and nouns, and (2) the output from the final layer. Although the data used in the current study are not manually crafted or controlled in any way similar to the above-mentioned studies, it attempts to add to the existing body of literature by (1) extending the analyses to all syntactic categories and (2) analyzing how BERT's predictions differ across layers. In light of all this, we ask the following questions:

1. Can the layer-wise linguistic knowledge found in structure studies be replicated with a behavior-based approach, namely, layer-wise masked word prediction analyses (§4.1.1)?
2. Do the results vary by syntactic category (§4.1.2)?

3 Experimental Setup

We used STREUSLE 4.4 (Schneider et al., 2018; Schneider and Smith, 2015), a corpus of web reviews written in English. This corpus contains 723 documents, 3,813 sentences, and 55,590 tokens in total with rich annotation of various syntactic and lexical-semantic information (e.g., annotation of 3,013 strong multiword expressions). The BERT's prediction data were prepared in the following way:

1. For each sentence, create n variants, where n is the number of tokens in the sentence, by replacing one token by [MASK] token.
2. For each variant (where one word is repalced with [MASK] in step 1) of each sentence, run vanilla BERT to generate a prediction from each layer $\ell \in L$.
3. For each of the n variants of each sentence,

where [MASK] is now replaced by a predicted token in step 2, POS-tag the predicted token to identify its syntactic category.

For the BERT model, we use `bert-base-uncased` because `bert-base` and `bert-large` have similar distributions of layers, which Rogers et al. (2020) call "stretch effect", although they do sometimes exhibit heterogeneous behaviors, such as responses to perturbation in word prediction (Ettinger, 2020). The model was retrieved from the PyTorch implementation of BERT by `huggingface` (Wolf et al., 2020).

For POS, the tag set of 17 POSs from Universal Dependencies (UD) v2 (Nivre et al., 2020) was used, and Stanza (Qi et al., 2020) was used for the automatic tagging of predicted tokens.

The above experiment resulted in the prediction of, and the tagging of, $L \times S \times N = 722{,}670$ masked tokens, where L, S, and N are the number of layers, the number of sentences, and the (mean) length of the sentences, respectively. In addition to analyzing the descriptive statistics, in order to quantify the relative contribution of each layer to POS match and word match, differential scores at each task (POS match or word match) for each layer $\Delta_T^{(\ell)}$ were obtained by computing the incremental gain from the previous layer (Equation 3 of Tenney et al., 2019a):

$$\Delta_T^{(\ell)} = Score_T^{(\ell)} - Score_T^{(\ell-1)} \tag{1}$$

As a summary statistic of these scores, (pseudo) expectation of differential scores (Equation 4 of Tenney et al., 2019a) was also calculated:

$$\bar{E}_\Delta[\ell] = \frac{\sum_{\ell=1}^{L} l \cdot \Delta_T^{(\ell)}}{\sum_{\ell=1}^{L} \Delta_T^{(\ell)}} \tag{2}$$

This is an "expected layer", at which the gain scores are centered around. If the differential scores were uniformly distributed, the expected layer would simply be the middle layer, which is layer 6. If the contribution of lower layers were higher (i.e., differential scores were higher at lower layers), then the expected layer would be lower than 6, and vice versa.

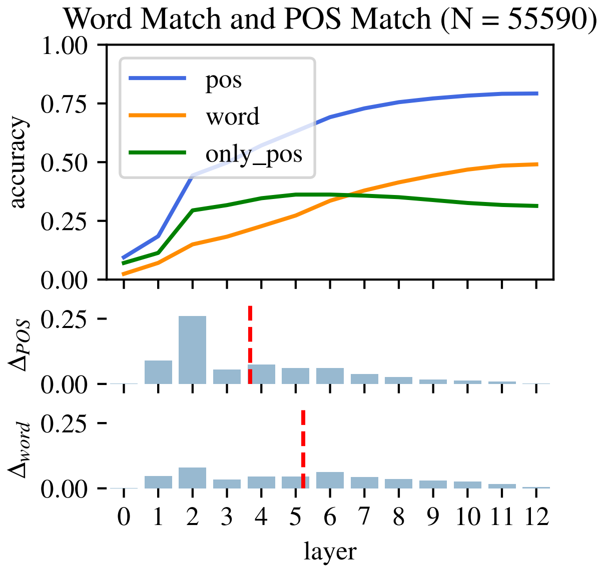

Figure 1: Layer-wise Accuracy of POS Match, Word Match, and POS Match without Word Match

4 Results

4.1 Quantitative Results

4.1.1 Overall

The top graph in Figure 1 illustrates the accuracy score[1] of POS match, word match, and POS match without word match (i.e., the predicted word is not the same as the original word, but is the same POS). Notably, POS match tends to increase at lower layers and approaches plateau towards the middle to high layers, whereas word match tends to increase linearly from lower to higher layers. Consequently, the proportion of the tokens with only POS match peaks at around layers 5 and 6 and starts declining beyond that point.

The middle and bottom graphs in Figure 1 illustrate the differential scores of POS match and word match, respectively. The vertical red dotted lines represent the expected layer defined in §3. The differential scores for POS match are clearly centered around lower layers followed by a sharp decline beyond middle layer, with the expected layer of 3.68. In contrast, the differential scores for word match are relatively more uniformly distributed across layers, and the expected layer is 5.22. This supports the findings from previous work that syntactic knowledge is more localizable at lower to middle

[1] Accuracy was chosen here for direct comparability; the proportion of the top predictions that are of the same POS as the original token, the proportion of the top predictions that are of the same word as the original token, and the proportion of the top predictions that are of the same POS but different word as the original token.

		N	POS	POS_M	word	word_M
open	ADJ	3169	4.04	4.24	6.45	6.35
	ADV	3080	3.42	3.76	5.74	5.30
	INTJ	108	3.48	**9.33**	7.13	**8.75**
	NOUN	7265	3.98	4.48	7.53	6.99
	PROPN	1406	**6.68**	6.11	**8.05**	7.88
	VERB	5328	3.96	3.68	6.73	6.38
closed	ADP	3368	3.16	3.52	5.01	5.18
	AUX	2950	3.10	4.43	5.14	5.08
	CCONJ	1803	5.48	5.32	5.88	4.74
	DET	3525	2.16	2.54	3.11	3.43
	NUM	555	5.70	6.81	6.73	7.23
	PART	1314	1.80	1.31	2.08	1.40
	PRON	5264	3.91	4.61	6.76	5.96
	SCONJ	808	5.05	4.45	5.71	5.45

Table 1: Expected Layer by UPOS. POS_M and word_M stand for POS_MWE and word_MWE, respectively.

layers (Tenney et al., 2019a; Liu et al., 2019) and that semantic knowledge is spread across layers (Tenney et al., 2019a).

4.1.2 By Syntactic Category

Table 1 summarizes the expected layers for POS match and word match for all tokens, as well as for tokens that are part of multiword expressions (MWEs), by UPOS.[2] In this section, we will focus on the former. First, in general, the expected layers for POS match and word match differ substantially by syntactic category. Whereas lower layers contribute much more for POS match for PART ($\bar{E}_\Delta[\ell]$ = 1.80), middle to higher layers contribute more for POS match for PROPN ($\bar{E}_\Delta[\ell]$ = 6.68). A similar observation is made for word match: on the one hand, lower layers contribute more for PART($\bar{E}_\Delta[\ell]$ = 2.08), and higher layers contribute more for PROPN ($\bar{E}_\Delta[\ell]$ = 8.05) on the other hand.

Although no straightforward generalizations can be made, for word match, we observe a tendency that expected layers tend to be higher when the original tokens are in open class, such as PROPN ($\bar{E}_\Delta[\ell]$ = 8.05) and NOUN ($\bar{E}_\Delta[\ell]$ = 7.53), whereas they tend to be lower when the original tokens are in closed class, such as PART ($\bar{E}_\Delta[\ell]$ = 2.08) and DET ($\bar{E}_\Delta[\ell]$ = 3.11).[3] This seems to suggest that higher layers tend to contribute more to word match for tokens in syntactic categories with more word types (i.e., open class), and that lower layers tend to contribute more for tokens in syntactic categories with fewer word types (i.e., closed class).

However, notable exceptions from closed class include NUM ($\bar{E}_\Delta[\ell]$ = 6.73) and PRON ($\bar{E}_\Delta[\ell]$ = 6.76). The former belongs to closed class because its atomic elements are finite (i.e., 0-9); however, with the infinite number of combinations of such elements, this class may be behaving similarly to open class. This is clearly not the case for the latter—PRON has a finite number of word types, which are fewer than the ones in open class. One plausible explanation is that identifying a correct pronoun requires a resolution of subject-verb agreement, which is shown to be handled well by BERT (Goldberg, 2019; van Schijndel et al., 2019) especially at layers 8 and 9 (Jawahar et al., 2019). However, upon closer examination, expected layers for personal pronouns in accusative case ($\bar{E}_\Delta[\ell]$ = 8.19) or those in (in)direct object positions ($\bar{E}_\Delta[\ell]$ = 8.08) are found to be much higher than those in nominative case ($\bar{E}_\Delta[\ell]$ = 6.34) or in subject positions ($\bar{E}_\Delta[\ell]$ = 6.29), although the latter should benefit from the subject-verb agreement resolution at higher layers. Given this observation, it may be the case that personal pronouns in accusative case or (in)direct object positions are more likely to necessitate long-distance coreference resolution in English, and such long-distance dependencies are shown to be handled better at higher layers (Jawahar et al., 2019). However, this hypothesis remains inconclusive (see §4.2 for more discussion).

4.1.3 Multiword Expressions

As an additional analysis of the effect of the number of potential candidates on expected layer, we calculate the expected layer only for tokens that are part of MWEs, based on the annotation of strong MWE in STREUSLE (Schneider et al., 2018; Schneider and Smith, 2015). Although the strong MWEs in STREUSLE consist of heterogeneous sets of expressions, such as idioms, light verb constructions, and noun compounds, we assume that, overall, this linguistic environment is more constrained and has fewer potential candidates for masked word prediction.

The POS_M and word_M columns in Table 1 represent the expected layers of POS match and word match only for the tokens that are part of MWEs, respectively. The expected layers are colored in red if they are higher for MWEs than for all tokens, and in blue if they are lower for MWEs than for all tokens. In general, on the one hand, for word match, they are lower for MWEs than for all tokens, which is congruent with the hypothesis from

[2] Miscellaneous tags, i.e. PUNCT, SYM, and X, are excluded from the analysis.

[3] Open and closed classes are based on the classification by UD project's (Nivre et al., 2020) website: `https://universaldependencies.org/u/pos/index.html`

ℓ	**prediction**
3	*to, a, him, me, them, her, the, us, one, people*
6	*him, me, her, them, us, to, people, everyone, it, you*
12	*us, her, me, we, them, everyone, our, him, it, stephanie*

Context: Stephanie's knowledge of the market and properties in our price range, made [MASK] (original: us) feel secure in our decision to buy when we did. (*reviews-341397-0002*)

Table 2: Selected Example 1 from STREUSLE

ℓ	**prediction**
3	*own, new, prison, personal, old, back, hospital, private, usual, current*
6	*own, private, bedroom, parking, damn, front, hotel, hospital, office, kitchen*
12	*car, garage, front, apartment, bedroom, office, cell, back, truck, elevator*

Context: they fixed my [MASK] (original: garage) doors in literally less than an hour. (*reviews-341397-0002*)

Table 3: Selected Example 2 from STREUSLE

§4.1.2 that lower layers contribute more when the number of potential candidates is relatively small. On the other hand, however, the expected layers for POS match tend to be higher for MWEs than for all tokens. The precise reason why this was the case is left for future work; however, we provide a potential account for these observations below.

One possible explanation for the higher expected layers of POS match is that semantic information plays an important role in predicting certain sequences of POSs observed in MWEs. For example, the common occurrences of noun compounds could be a contributing factor to the higher expected layer for POS match only for NOUNs that are part of MWE ($\bar{E}_\Delta[\ell] = 4.48$) compared to that of all NOUNs ($\bar{E}_\Delta[\ell] = 3.98$). Given that the meaning of the second token (the head of the compound) is crucial in detecting its (dis)preference on forming a compound, it may require more semantic information for BERT to correctly identify that the first token is NOUN rather than ADJ, resulting in a higher expected layer. Indeed, for all NOUNs that are part of MWE, the most common incorrect prediction was ADJ at all layers from layer 2 through 12, which was not the case for NOUNs that are not part of MWE (see §4.2 for an example).

4.2 Qualitative Results

In this section, we present a set of selected examples from the STREUSLE corpus to illustrate the observations made in §4.1.

Table 2 illustrates the identification of a personal pronoun at each layer of BERT (only showing layers 3, 6, and 9). From lower to higher layers, it is clear that the ranking of the correct pronoun *us* is steadily promoted. In fact, it is not until layer 11 that the correct pronoun *us* receives the highest prediction probability. In §4.1.2, one hypothesis that can potentially account for the higher expected layer of PRON (personal pronouns in object positions or in accusative case in particular) was the long-distance dependency. In Table 2, pronouns *our* and *we* are readily available in relatively close proximity, but the correct pronoun *us* is not identified until layer 11. This seems to suggest that pronouns that are ACC-marked or in object positions pose unique challenges not explicable only by the distance of the dependency.

Table 3 illustrates BERT's predictions of the first token of a noun compound *garage doors*. As discussed in §4.1.3, at layer 3, many of the predictions are generic adjectives (e.g., *own, new, old, private, usual, current*), although the meaning of the word *door* seems to be captured to some extent, as we can see from some of the predictions (e.g., *prison, back, hospital*). At layer 6, such prediction of nouns that are specific to the meaning of the word *door* becomes more dominant. This is even more so at layer 12, where such nouns occupy most of the predictions despite the presence of a cue, *my*, which strongly collocates with *own*. This supports our observation that, for some syntactic categories including NOUN, MWE's production of certain sequences of POSs necessitates more semantic information to restore the POS of the original word, resulting in a higher expected layer.

5 Conclusion

In this study, we set out to investigate if (1) the layer-wise linguistic knowledge found in structure studies can be replicated with a behavior-based design and if (2) the results vary by syntactic category. By analyzing BERT's layer-wise masked word prediction, we have shown that the localization of linguistic knowledge found in various probing studies was indeed replicated; more specifically, syntactic knowledge was encoded primarily in lower layers, whereas semantic knowledge was spread across the 12 layers.

We also observed that the contribution of particular layers on syntactic and semantic information varied substantially, depending on the syntactic category (i.e., UPOS) and on the syntactic class (i.e., open vs. closed class) more generally, of the original token. Hypothesizing that the number of potential candidates is one of the contributing factors to this difference, we showed that, in general, the expected layers were higher for POS match and lower for word match for the tokens that are part of MWEs (a supposedly more constrained environment).

Our contribution is twofold. First, by leveraging BERT's layer-wise outputs, we confirmed the previous studies without relying on external probing classifiers or introducing extra parameters that can potentially obfuscate the locus of the observed linguistic knowledge (i.e., language model vs. probing classifier). Second, by extending the analyses to all open and closed class categories rather than limiting the scope to popular content-words, such as verb, noun, and adjective, we show that the encoding of syntactic and semantic knowledge about words of different UPOS varies substantially.

Lastly, we acknowledge that this study has a few limitations. First, the layer-wise masked word prediction essentially feeds intermediate layers directly to the classification layer, thereby inferring the linguistic information encoded in the intermediate layers. However, this is not what BERT is trained for; that is to say, arguably, only the final layer is optimized for the masked word prediction task, and other layers are not. Hence, the intermediate layers' lower POS and word match accuracy may not be due to the "absence" of syntactic or semantic knowledge encoded in those layers; rather, they may simply suggest that those intermediate layers are not trained for such tasks.

Second, although we provided a possible explanation for our observations and showed a few examples that seem to support our hypotheses, these are highly speculative and not meant to prove anything. We consider this a limitation of our approach, and a more controlled experiment is needed to make stronger claims or to test our hypotheses, and this is left for future work.

Acknowledgements

We thank the three anonymous reviewers for their insightful and constructive feedback. We also thank Dr. Ellie Pavlick and others at Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL) for their helpful comments.

References

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. 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.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Luke Gessler and Nathan Schneider. 2021. BERT has uncommon sense: Similarity ranking for word sense BERTology. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 539–547, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

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.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. 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.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. 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 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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.

Yichu Zhou and Vivek Srikumar. 2021. DirectProbe: Studying representations without classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5070–5083, Online. Association for Computational Linguistics.