Can We Use Probing to Better Understand Fine-Tuning and Knowledge Distillation of the BERT NLU?

Jakub Hościłowicz[1,2][a], Marcin Sowański[1,2][b], Piotr Czubowski[1][c] and Artur Janicki[2][d]

[1] *Samsung R&D Poland Institute, Warsaw, Poland*

[2] *Warsaw University of Technology, Warsaw, Poland*

Keywords: Probing, Natural Language Understanding, Dialogue Agents, BERT, Fine-Tuning, Knowledge Distillation.

Abstract: In this article, we use probing to investigate phenomena that occur during fine-tuning and knowledge distillation of a BERT-based natural language understanding (NLU) model. Our ultimate purpose was to use probing to better understand practical production problems and consequently to build better NLU models. We designed experiments to see how fine-tuning changes the linguistic capabilities of BERT, what the optimal size of the fine-tuning dataset is, and what amount of information is contained in a distilled NLU based on a tiny Transformer. The results of the experiments show that the probing paradigm in its current form is not well suited to answer such questions. Structural, Edge and Conditional probes do not take into account how easy it is to decode probed information. Consequently, we conclude that quantification of information decodability is critical for many practical applications of the probing paradigm.

1 INTRODUCTION

In recent years significant progress has been made in the natural language processing (NLP) field. Foundation models, such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), remarkably pushed numerous crucial NLP benchmarks and the quality of dialogue systems. Current efforts in the field often assume that progress in foundation models can be made by either increasing the size of a model and dataset or by introducing new training tasks. While we acknowledge that this approach can lead to improvements in many downstream NLP tasks (including natural language understanding – NLU), we argue that current model evaluation procedures are insufficient to determine whether neural networks "understand the meaning" or whether they simply memorize training data. Knowing this would allow us to build better NLU models.

In the last few decades, measuring the knowledge of neural networks has been one of key challenges for NLP. Perhaps the most straightforward way of doing this is to ask models commonsense questions and measure how often their answer is correct (Bisk et al., 2020). However, as noted in Bender and Koller (2020), such an approach might be insufficient to prove whether a neural network "understands meaning" because the answers given by foundation models might be the result of memorization of training data and statistical patterns. And since this approach is not conclusive, a new set of methods that focus on explaining the internal workings of neural networks has been proposed (Limisiewicz and Marecek, 2020; Clark et al., 2019). Among the most prominent are probing models, which are usually used to estimate the amount of the various types of linguistic information contained in neural network layers (Hewitt and Manning, 2019; Tenney et al., 2019). In the probing paradigm, a simple neural network is applied to solve an auxiliary task (e.g., part of speech tagging) to a frozen model. The assumption is that the higher the performance of such a probing neural network, the higher the amount of the respective type of information a probed model contains.

Many efforts have been made to better understand the probing paradigm, its purposes, and its usefulness in generating valuable insights from studies on neural networks. Nevertheless, it is still unclear how exactly probing should be used in the NLP field and what conclusions can be drawn from probing results (Ivanova et al., 2021). Probing methods are inspired by neu-

[a] https://orcid.org/0000-0001-8484-1701
[b] https://orcid.org/0000-0002-9360-1395
[c] https://orcid.org/0000-0002-1088-6154
[d] https://orcid.org/0000-0002-9937-4402

Hościłowicz, J., Sowański, M., Czubowski, P. and Janicki, A.
Can We Use Probing to Better Understand Fine-Tuning and Knowledge Distillation of the BERT NLU?.
DOI: 10.5220/0011724900003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 625-632
ISBN: 978-989-758-623-1; ISSN: 2184-433X

roscience where an analogical approach is used to better understand how information is processed in a human brain (Glaser et al., 2020). Most importantly, Kriegeskorte and Douglas (2019) conclude that a decoding probe model can only reveal whether a particular type of information is present in a certain brain region. Information perspective is also present in NLP – Pimentel et al. (2020) state that one should always select the highest-performing probe, because it gives the best estimation of the lower bound of information amount in a neural network.

Our research is motivated by the questions we faced while working on NLU models used in dialogue agents. Development of production-ready BERT-based NLU models requires deep understanding of phenomena that occur during fine-tuning and knowledge distillation (KD). Especially important issues are "catastrophic forgetting" and choice of optimal fine-tuning dataset size. Probing seems to be a relevant tool to analyze these issues, but in this work we will show that its practical usability in the NLU context is low.

Our article is structured as follows. First, in Section 2, we present a review of related work in the area. Next, in Sections 3 and 4, we describe the design and results of our experiments. In Section 5, we discuss the results of our experiments in the context of related works from both machine learning and neuroscience. Finally, in Section 6, we summarize our arguments and outline a proposed solution.

2 RELATED WORK

In the literature, three types of probing are usually described. *Structural probing*, introduced by Hewitt and Manning (2019), which conceptually aims at estimating the amount of syntactic information through reconstruction of dependency trees from vector representations returned by neural networks. *Edge probing* (Tenney et al., 2019) includes a wide array of subsentence NLP tasks. All the tasks can be formulated as syntactic or semantic relations between words or sentences. *Conditional probing* (Hewitt et al., 2021) is a framework which can be applied to any probing method (e.g., conditional structural probing). Conceptually, conditional probing tells how much information is present in a given layer of a model, provided that this information is not present in a chosen baseline (e.g., embedding layer). Consequently, conditional probing allows for better comparison of models with different architectures – it grounds probing results in non-contextual embedding layers of models.

Some authors used probing to better understand how linguistic information is acquired by models and how it changes during fine-tuning (Kirkpatrick et al., 2017; Hu et al., 2020). Durrani et al. (2021) and Pérez-Mayos et al. (2021) analyze the phenomenon of "catastrophic forgetting" and generalization in the NLU context. Zhang et al. (2021) and Pérez-Mayos et al. (2021) used probing as a tool to estimate the optimal size of pretraining data for NLU tasks. Zhu et al. (2022) concluded that probing is an efficient tool for that, because it does not require gradient updates of the entire model. Probing has also been used to compare amounts of linguistic knowledge in neural networks (Nikoulina et al., 2021; Liu et al., 2019).

The majority of publications related to BERT fine-tuning report a drop in probing results after fine-tuning. Durrani et al. (2021) conclude that the decrease in probing results means that fine-tuning leads to catastrophic forgetting. Mosbach et al. (2020) hypothesize that after fine-tuning, linguistic information might be less linearly separable and hence harder to detect for a probing model. The broad fine-tuning analysis of Merchant et al. (2020) (including probing models) guides authors towards the hypothesis that fine-tuning does not introduce arbitrary (negative) changes to a model's representation, but mainly adjusts it to a downstream task. Relying on improvements and new insights in the probing field (e.g., conditional probing), we wanted to continue along this path in order to understand what happens in our particular NLU scenario.

Probing in the KD context was used by Kuncoro et al. (2019). The authors concluded that their new neural network architecture has better syntactic competencies than the baseline. Similarly, Fei et al. (2020) used probing to assess the language competencies of a distilled student model. The fact that the probing results of the student model are close to those of the teacher's leads the authors to the conclusion that the student is effective at capturing syntax. In a similar manner, we wanted to use probing to measure the linguistic capabilities of a distilled NLU model.

3 METHODS AND MODELS

To better illustrate our interpretation of probing, we introduced the following terminology (see also Figure 1): we define the Primary Decoder as the decoder that was used to train or fine-tune a given Encoder (e.g., BERT). For example, in the case of our BERT fine-tuning, it is the NLU head. The Secondary Decoder is an external decoder that was not used for training or fine-tuning of the Encoder (e.g., probing decoder). All probing results noted in this publication are the results of the last layer of the Encoder (e.g., the last layer of BERT).

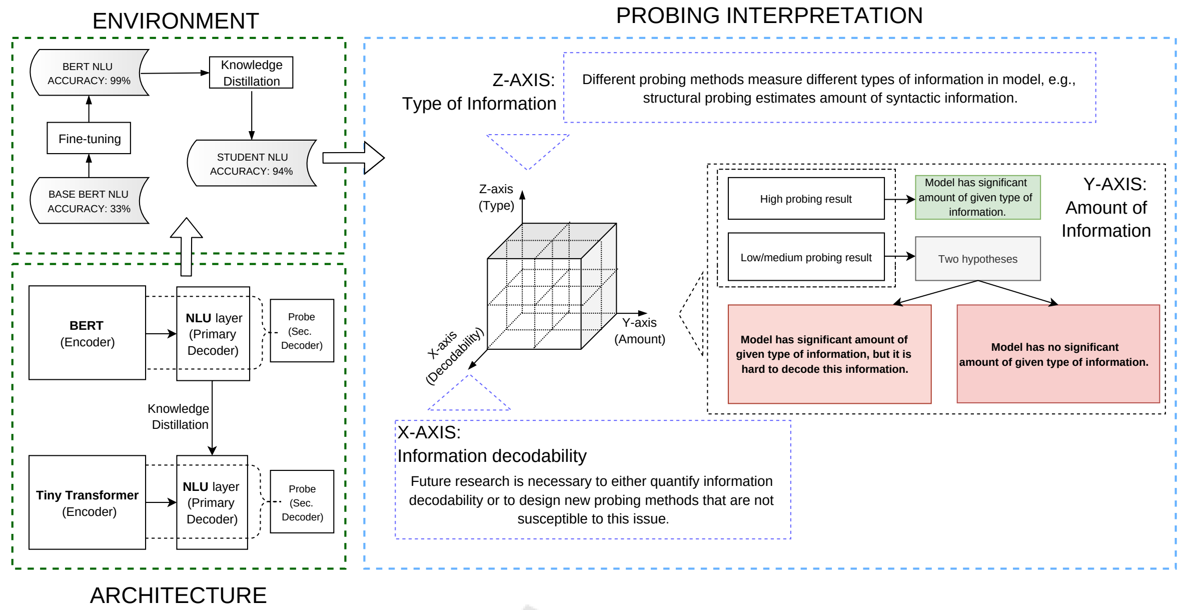
ENVIRONMENT
BERT NLU ACCURACY: 99%
Knowledge Distillation
Fine-tuning
STUDENT NLU ACCURACY: 94%
BASE BERT NLU ACCURACY: 33%
BERT (Encoder)
NLU layer (Primary Decoder)
Probe (Sec. Decoder)
Knowledge Distillation
Tiny Transformer (Encoder)
NLU layer (Primary Decoder)
Probe (Sec. Decoder)
ARCHITECTURE
PROBING INTERPRETATION
Z-AXIS: Type of Information
Different probing methods measure different types of information in model, e.g., structural probing estimates amount of syntactic information.
Z-axis (Type)
Y-axis (Amount)
X-axis (Decodability)
High probing result
Model has significant amount of given type of information.
Low/medium probing result
Two hypotheses
Y-AXIS: Amount of Information
Model has significant amount of given type of information, but it is hard to decode this information.
Model has no significant amount of given type of information.
X-AXIS: Information decodability
Future research is necessary to either quantify information decodability or to design new probing methods that are not susceptible to this issue.

Figure 1: Probing interpretation in typical NLU scenario. The "Architecture" sub-graph describes modeling design and terminology. "Environment" gives an overview of the pipeline. Finally, the "Probing Interpretation" sub-graph describes how probing can be interpreted and what limitations we see.

3.1 Probing Methods Used

We investigated three distinct types of probing methods: structural, edge and conditional. The performance of the structural probing model is measured with Unlabeled Undirected Attachment Score (UUAS) – which represents the percentage of undirected edges placed correctly against the gold syntactic tree. We used the part-of-speech (POS) tagging-based variant of edge probing – its metric is F1 score. As a baseline for conditional probing, we used the non-contextual embedding layer of the probed neural network, similarly to Hewitt et al. (2021). The metric for conditional probing is determined by the chosen probing method, e.g., for conditional structural probing it is UUAS.

3.2 Joint NLU

There are many NLP tasks that are considered part of the NLU field. In this paper, we define NLU in the way which is most popular in the dialogue agent context – as the intent classification and slot-filling tasks (Weld et al., 2021). An intent represents a command uttered to a dialogue system and slots are parameters of that command. For example, in the utterance "play Radiohead on Spotify", 'Radiohead' and 'Spotify' are slots and 'to_play_music' is an intent.

As an NLU architecture, we used Joint BERT (Chen et al., 2019). This NLU model is created by extending BERT with two softmax classifiers corresponding to intents and slots respectively. To measure NLU quality, we used semantic frame accuracy. It is calculated as a fraction of test sentences where both intent and all slots were correctly predicted. We tested three variants of the NLU architecture, differentiated by the way in which BERT output vectors are passed to the intent classification layer:

- **Pooled IC**, where input to IC is a vector corresponding to BERT's special [CLS] token. This is the approach used in (Chen et al., 2019),
- **Average IC**, where input to IC is average of BERT's output vectors,
- **Sum IC**, where input to IC is sum of BERT's output vectors,

3.3 Knowledge Distillation for Joint NLU

The purpose of KD is to train the student model through imitation of the teacher model. The objective is to minimize KD Loss L_{KD}, which measures divergence between student, training data and teacher. Because Joint NLU (Chen et al., 2019) is based on a multitask paradigm, divergence losses of two tasks, intent IC classification and slot filling SC, are minimized simultaneously ($L_{KD} = L_{IC} + L_{SC}$). L_{IC} is analogical to:

$$L_{SC} = H(y, \sigma(z_s)) + H(\sigma(z_t; \rho), \sigma(z_s; \rho))$$

where H is cross entropy loss function, y are ground truth slot labels, z_s and z_t are student and teacher logits respectively. Additionally, softmax function σ parametrized by temperature ρ, is applied to logits.

Distillation was performed to a randomly initialized student model with transformer architecture analogical to BERT, but reduced to two layers and two attention heads. BERT NLU (Pooled) was used as the teacher model. As a point of reference, we also included the results of a tiny model with the same architecture as student, but trained without distillation (Tiny Transformer NLU).

3.4 Data

We used a subset of Leyzer (Sowański and Janicki, 2020) that consisted of five domains with 51 intents and 29 slots. We annotated Leyzer with the Stanford Dependency Parser (Chen and Manning, 2014) so that it could be used in the probing context. Additionally, for probing analysis, we used the Universal Dependencies (UD) corpus (Nivre et al., 2020), which is manually annotated and consists of general-type sentences unrelated to NLU. The Leyzer subset we used contains 5200 sentences and the UD corpus consists of 16,621 in total. We used an 80%, 10%, 10% train/test/validation split. To better measure the generalization power of models, we manually constructed a small (164 test cases) Malicious NLU testset. It is based on the Leyzer testset, but consists of sentences which were designed to better measure the generalization power of NLU models. Such test cases contain named entities and grammatical constructs not present in the training set.

4 EXPERIMENTAL RESULTS

4.1 Fine-Tuning Analysis

To get a better perspective on the analyzed phenomena, we evaluated NLU in two variants. In the case of BERT NLU, the BERT model is fine-tuned with NLU head. In the Frozen BERT variant, BERT's weights were fixed during NLU decoder training. Consequently, Frozen BERT gives baseline results (BERT is not fine-tuned). We reported probing results both on Leyzer and UD to give perspective on how in-domain (Leyzer) probing results differ from out-of-domain (UD).

Detailed results presented in Table 1 show that, as expected, fine-tuning improved NLU accuracy during the course of training. An inverse trend can be observed in relation to probing results. UUAS gets lower as fine-tuning progresses. In the end, fine-tuned NLU model probing results were significantly lower than those of Frozen BERT. As presented in Figure 2, exactly the same tendencies were observed when we gradually increased the size of the fine-tuning dataset.

Table 1: Results of structural probing on BERT NLU model tested on Leyzer and Universal Dependency (UD) testsets. Linear baseline based on Hewitt and Manning (2019).

Model Variant	Epoch	NLU Accuracy	Structural probing (Leyzer)	Structural probing (UD)
Linear baseline	-	-	-	0.49
Frozen BERT	100	0.33	0.82	0.66
BERT NLU (Pooled)	1	0.0	0.79	0.60
	5	0.45	0.75	0.56
	10	0.85	0.70	0.55
	30	0.97	0.55	0.52
	60	0.97	0.58	0.52
	100	0.97	0.50	0.50

In Figure 3, we present the results of the experiment with three different NLU architectures (as described in subsection 3.2). At the end of the fine-tuning process, the NLU results were nearly the same, while the UUAS differed significantly for each architecture. A strong downward trend is visible in all possible cases. All tendencies from this chapter were the same in the case of conditional probing; see example in Table 2.

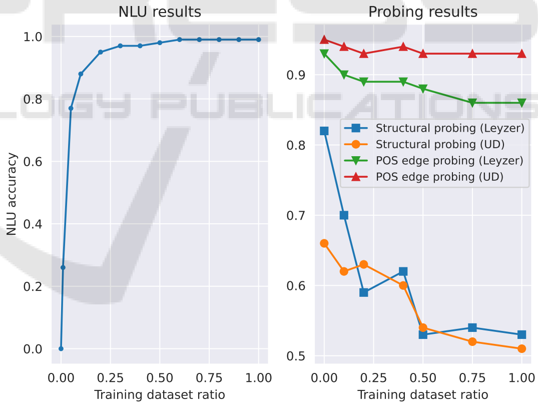

Figure 2: Influence of dataset size on NLU accuracy and probing results.

4.2 Probing Knowledge Distillation

The first observation drawn from the NLU results presented in Table 2 is that both Student and Tiny Transformer have lower generalization power than BERT NLU. If we compare Student NLU to its non-distilled version, we observe non-negligible test accuracy gain resulting from the KD process. In our case, the temperature was a key factor for distillation – the best NLU result is achieved for the student distilled with

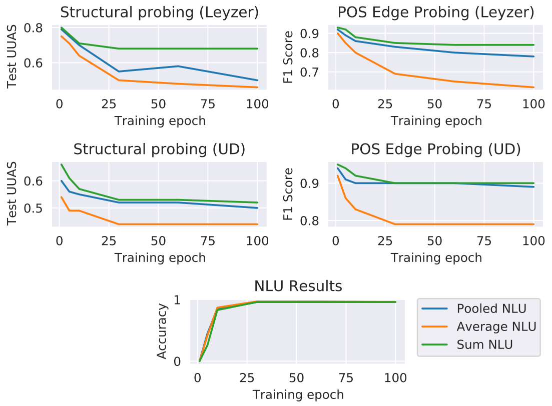

Figure 3: Structural and edge probing results in fine-tuning scenario on Leyzer and UD datasets, for three variants of NLU architecture.

$T = 0.01$ (presented in Table 2). Neither temperature nor choice of teacher significantly influences the probing results of the Student NLU.

In Table 2 we focused only on conditional probing because it gives a more valuable comparison of models with different architectures; see also Section 2. Both variants of conditional probes indicate that BERT NLU, Student NLU, and Tiny Transformer NLU have a near-zero amount of syntactic and part-of-speech information in the last layers of their Encoders. Specifically, if we measure only information not available in the respective non-contextual baselines, then the amount of information in the last Encoders' layers for all mentioned models is the same and close to zero. We refrain from comparing teacher and student using UD corpus because our student is not pre-trained on any kind of general corpus. Nonetheless, UD results give a scale for conditional probing results. For Frozen BERT, the result on UD corpus is 0.15 for conditional structural probing and 0.11 for conditional edge probing.

Table 2: Results for knowledge distillation and probing.

	NLU acc.	Malicious NLU acc.	Conditional Structural Probing (Leyzer)	Conditional Edge Probing (Leyzer)
Frozen BERT	0.33	0.10	0.05	0.07
BERT NLU (Pooled)	0.97	0.56	0.02	0.01
Student NLU	0.94	0.27	0.01	0.01
Tiny Transformer NLU	0.90	0.23	0.01	0.01

5 DISCUSSION

The significant decrease in probing results on a general UD corpus can be explained by a known phenomenon: when we fine-tune BERT on an NLU task, its general linguistic knowledge is destroyed ("catastrophic forgetting"). However, the decrease in probing results on the Leyzer corpus is not so easy to interpret. BERT's fine-tuning on Leyzer leads to substantial deterioration of probing results on precisely the same dataset. We initially presumed that fine-tuning on Leyzer would increase related linguistic knowledge, which would be reflected in improvement in the probing results. However, both experiments suggest that fine-tuning optimizes downstream task accuracy, which comes at the expense of the linguistic knowledge associated with Leyzer.

The experiments with different NLU decoder architectures show significant changes in probing results, while at the same time NLU accuracy was not affected that much. Depending on the NLU decoder mode (pooled, averaged, sum), edge and structural probing results can vary from around 0.45 to 0.70 UUAS and 65% to 85% F1 Score. This observation suggests that the NLU decoder can play an important role in how linguistic information in the Encoder is structured.

In the conditional framework, the downward probing tendencies of BERT's fine-tuning do not change. Conditional results on Leyzer are nearly the same for teacher and non-pretrained student NLUs. Both models achieve near-zero conditional probing results and if we apply a standard interpretation, we can conclude that neither contains a significant amount of syntactic and part-of-speech information in the last layers of their Encoders (compared to the respective non-contextual baselines).

All mentioned observations guided us toward the two main hypotheses presented in "Amount of Information" in Figure 1. The deterioration of probing results might mean either that the amount of linguistic information decreased or that it is harder to decode for a probing model.

The purpose of the Primary NLU Decoder is to structure BERT's knowledge so that it can effectively extract it and achieve high accuracy on a downstream task. The Primary Decoder does not know about the existence of Secondary Decoders and its goal is not to make information in the Encoder easily extractable for them, but to reduce NLU training loss. Consequently, making definitive conclusions ("fine-tuning leads to catastrophic forgetting") relying on probing results could be misleading. The degree of "catastrophic forgetting" (as measured with probing) could be much smaller than we think, or even non-existent. However, as shown in Figure 1, without inclusion of the decodability issue in the probing paradigm, such considerations remain inconclusive. Low probing results can mean either that information is not present, or that it is hard to decode.

6 CONCLUSIONS AND FUTURE WORK

Relying on insights from the experiments, neuroscience (Kriegeskorte and Douglas, 2019; Ivanova et al., 2021) and NLP, we conclude that probing has small usability for analysis of NLU models. Current probing methods do not consider how easy it is to decode probed information, hence they only give an estimation of the lower bound of information amount (Pimentel et al., 2020). Future research is necessary to either quantify information decodability (as visualized in Figure 1) or to design new probing methods that are not susceptible to this issue.

However, to measure decodability, firstly, information in neural networks must be defined in a rigorous way. Ultimately, similar to (Tschannen et al., 2020), we conclude that a new concept of information is important for the future of NLP research. One approach to how information can be defined in neural networks is presented by Xu et al. (2020), but we have not yet found any works about information decodability.

Another path is to use probing in a neuroscience manner, as proposed by Kriegeskorte and Douglas (2019). Instead of trying to answer the question “How much information of a given type is in the neural network?”, we can focus on the question “Is a given type of information present in the neural network?” Consequently, we use probing results as a component of p-value for the hypothesis that there is no significant amount of a given type of information in a given neural network layer. This path implies that we should focus more on new types of information which are less obvious than those currently probed. Information decodability also constitutes an issue for this approach. However, in our opinion, with proper design of hypothesis testing (including heuristics about decodability of information), this approach can give more reasonable insights than the “How much information” paradigm.

To summarize, the main contributions of our work are as follows:

- We showed that current probing methods are of low usability for analysis of NLU models. Without careful interpretation, they might lead to wrong conclusions.
- We presented a clear interpretation of probing in the NLP context. This interpretation implies that information decodability is a large obstacle for many practical applications of probing methods.

REFERENCES

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. ACL.

Bisk, Y., Zellers, R., Gao, J., Choi, Y., et al. (2020). Piqa: Reasoning about physical commonsense in natural language. In *Proc. AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. ACL.

Chen, Q., Zhuo, Z., and Wang, W. (2019). BERT for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of bert's attention. In Linzen, T., Chrupala, G., Belinkov, Y., and Hupkes, D., editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Durrani, N., Sajjad, H., and Dalvi, F. (2021). How transfer learning impacts linguistic knowledge in deep NLP models? In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4947–4957. Association for Computational Linguistics.

Fei, H., Ren, Y., and Ji, D. (2020). Mimic and conquer: Heterogeneous tree structure distillation for syntactic NLP. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 183–193. Association for Computational Linguistics.

Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., and Kording, K. P. (2020). Machine learning for neural decoding. *Eneuro*, 7(4).

Hewitt, J., Ethayarajh, K., Liang, P., and Manning, C. D. (2021). Conditional probing: measuring usable information beyond a baseline. In Moens, M., Huang, X., Specia, L., and Yih, S. W., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1626–1639. Association for Computational Linguistics.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1*, pages 4129–4138.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1725–1744. Association for Computational Linguistics.

Ivanova, A. A., Hewitt, J., and Zaslavsky, N. (2021). Probing artificial neural networks: insights from neuroscience. *CoRR*, abs/2104.08197.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proc. of the National Academy of Sciences*, 114(13):3521–3526.

Kriegeskorte, N. and Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current opinion in neurobiology*, 55:167–179.

Kuncoro, A., Dyer, C., Rimell, L., Clark, S., and Blunsom, P. (2019). Scalable syntax-aware language models using knowledge distillation. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3472–3484. Association for Computational Linguistics.

Limisiewicz, T. and Marecek, D. (2020). Syntax representation in word embeddings and neural networks - A survey. In Holena, M., Horváth, T., Kelemenová, A., Mráz, F., Pardubská, D., Plátek, M., and Sosík, P., editors, *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020), Hotel Tyrapol, Oravská Lesná, Slovakia, September 18-22, 2020*, volume 2718 of *CEUR Workshop Proceedings*, pages 40–50. CEUR-WS.org.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1073–1094. Association for Computational Linguistics.

Merchant, A., Rahimtoroghi, E., Pavlick, E., and Tenney, I. (2020). What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Mosbach, M., Khokhlova, A., Hedderich, M. A., and Klakow, D. (2020). On the interplay between finetuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In Alishahi, A., Belinkov, Y., Chrupala, G., Hupkes, D., Pinter, Y., and Sajjad, H., editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pages 68–82. Association for Computational Linguistics.

Nikoulina, V., Tezekbayev, M., Kozhakhmet, N., Babazhanova, M., Gallé, M., and Assylbekov, Z. (2021). The rediscovery hypothesis: Language models need to meet linguistics. *Journal of Artificial Intelligence Research*, 72:1343–1384.

Nivre, J., de Marneffe, M., Ginter, F., Hajic, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F. M., and Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4034–4043. European Language Resources Association.

Pérez-Mayos, L., Ballesteros, M., and Wanner, L. (2021). How much pretraining data do language models need to learn syntax? In Moens, M., Huang, X., Specia, L., and Yih, S. W., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1571–1582. Association for Computational Linguistics.

Pimentel, T., Valvoda, J., Maudslay, R. H., Zmigrod, R., Williams, A., and Cotterell, R. (2020). Information-theoretic probing for linguistic structure. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4609–4622. Association for Computational Linguistics.

Sowański, M. and Janicki, A. (2020). Leyzer: A dataset for multilingual virtual assistants. In Sojka, P., Kopeček, I., Pala, K., and Horák, A., editors, *Proc. Conference on Text, Speech, and Dialogue (TSD 2020)*, pages 477–486, Brno, Czechia. Springer International Publishing.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., and Pavlick, E. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. (2020). On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Weld, H., Huang, X., Long, S., Poon, J., and Han, S. C. (2021). A survey of joint intent detection and slot-filling models in natural language understanding. *arXiv preprint arXiv:2101.08091*.

Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. (2020). A theory of usable information under computational constraints. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhang, Y., Warstadt, A., Li, X., and Bowman, S. R. (2021). When do you need billions of words of pretraining data? In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1112–1125. Association for Computational Linguistics.

Zhu, Z., Shahtalebi, S., and Rudzicz, F. (2022). Predicting fine-tuning performance with probing. *arXiv preprint arXiv:2210.07352*.