NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Ksenia E. Chistyakova, Tatiana B. Kazakova

GRAMMAR IN LANGUAGE MODELS: BERT STUDY

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS

WP BRP 115/LNG/2023

This Working Paper is an output of a research project implemented within NRU HSE's Annual Thematic Plan for Basic and Applied Research. Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

Ksenia E. Chistyakova[1], *Tatiana B. Kazakova*[2]

GRAMMAR IN LANGUAGE MODELS: BERT STUDY[3]

The problem of language models' interpretation is extensively inspected, but no universal answers have been found. Our study offers to combine widely accepted probing methods with a novel approach to a neural network under investigation. We propose to break grammatical forms on the pre-training step in order to get two "sibling" models, as it casts some light on how different linguistic features are encoded and distributed across the neural language architecture.

Project repository: https://github.com/skipdividedd/bert_grammar

Keywords: probing, language models, transformers, BERT.

JEL Classification: Z

[1] HSE University. E-mail: ksevgch@gmail.com
[2] HSE University. Laboratory for Arctic Social Sciences and Humanities. E-mail: tanusha.kazakova@gmail.com
[3] This paper was prepared within the framework of the HSE University Basic Research Program in 2023.

1 Introduction

While deep neural networks have been on the rise, it is still unclear how they capture and store knowledge about both the world and the language from the perspective of human interpretation. Although a model's architecture has been explained in detail (Vaswani et al., 2017), its internal workings remain "a black box".

Many researchers have approached this problem, trying to analyze models' representations. A number of experiments have been previously conducted, such as amnesic probing (Elazar et al., 2021), layer-wise probing (Fayyaz et al., 2021), chronological probing (Voloshina et al., 2022). Still, a common and widely used strategy is to train a linear classifier to predict labels of a linguistic category, treating the activations generated from the neural network as features (Belinkov, 2022). If the probe shows good performance, it might be legitimate to say that the model encodes the target property somehow. Nonetheless, it is still unknown where that knowledge is stored.

This paper aims to offer a technique, which seeks to get more insight into the inner mechanisms of how deep neural networks actually work. We formulate our method as follows:

• choose a neural network architecture;

• make two datasets in your target language which contain practically the same sentences, but differ in a way that some predefined grammatical forms are randomly changed ("spoiled"), e.g. the grammatical number of noun;

• train two models (we further address them as "good" and "broken" models);

• choose target linguistic properties under probing investigation;

• train probes to predict a grammatical feature using the models' representations;

• conduct further comparisons of the "sibling" pre-trained models and of probing performance, do neuron-level analysis.

We prove this approach to be promising as long as it enables to get a better understanding of how specific information is encoded inside a deep neural network.

2 Related work

There is a "zoo" of probing methods, but generally the proposed ones may be called either correlational or causal. While correlational probes indicate that the model has the knowledge about the target linguistic property, the purpose of causal techniques is to determine whether the model uses the information discovered by a probe.

2.1 Correlational Probing

The methodology is quite straightforward: having a dataset of sentence - label pairs, a classifier – a probe – is trained on the latent neural network model representations, so that the desired information (label) is predicted. High prediction score is regarded to be an indicator that such property can be extracted from the activations as it is actually encoded there.

This approach is used in the study done by Serikov et al. (2022), where the authors probe mBERT and XLM-RoBERTa models for 104 languages and 80 morphosyntactic features with both linear (Logistic Regression) and nonlinear (MLP) classifiers.

Voloshina et al. (2022) do a chronological probing investigation to show the changes within the language model during pre-training step, making use of logistic regression on top of embeddings of mBERT and T5.

Still, the method is critised (Belinkov, 2022). It is highlighted that the probe is "a proxy", which raises the issue of whether the classifier actually reflects some correlations instead of just memorizing patterns on its own.

The problem has been addressed through control tasks and selectivity criterion (Hewitt and Liang, 2019). It is defined as the difference between accuracy of probes trained on "true" data and one with randomized (shuffled) labels. The higher selectivity, the better performance of the classifier, the more we can trust the results.

2.2 Causal Probing

Prior work in this field focuses on interventions, either modifying a model’s representations or input data, in order to get so-called representational and templated / naturalistic "counterfactuals" respectively as the result. Such an approach is devised to evaluate the importance of a specific type of information in its relevance to some linguistic tasks.

Amnesic probing with INLP (iterative null space projection) was presented by Elazar et al. (2021), so that the authors erased information about POS (part-of-speech) from BERT's activations (Devlin et al., 2019) and then fed them to the word prediction layer to measure the performance after representational interventions.

Finlayson et al. (2021) studied syntactic agreement mechanisms in neural language models. The researchers created templated counterfactuals – pairs of sentences, where a single analyzed property is changed, e.g. "The athlete confuses/*confuse".

Naturalistic counterfactuals are widely used to solve the problem of eliminating gender bias (Zmigrod et al., 2019). The key distinction from templates is that dependency structure of the sentence is taken into account. Naturalistic counterfactuals are also used for linguistic probing (Amini et al., 2023).

In our work we take advantage of both causal and correlational techniques, as we generate templated counterfactuals for BERT pre-training and then use a probing classifier to study linguistic properties.

3 Methodology

In our work we lean heavily on the NeuroX framework (Dalvi et al., 2019b), which is a toolkit for neuron-level analysis (Dalvi et al., 2019a). It makes use of a logistic regression classifier with elastic-net regularization and enables to select N% of top neurons for a specific linguistic task based on the weights of the trained probe. Then it is possible to train a probe only on a subset of neurons for the same task. There is also an option to conduct a general probing experiment layer-wise.

Our experiment is largely premised on the toolkit. We also introduce control task, which was mentioned previously, to get more solid and trustful results.

As there is a vast body of literature on BERT and its representations, e.g. Tenney et al. (2019) and Rogers et al. (2020), we have chosen it to be a base model under investigation.

4 Data

4.1 Pre-training Data

As we exploit the previously discussed idea of templates, we propose to change values in some chosen grammatical categories in a sentence not for probing, but for a neural language model pre-training, so that it "breaks" in its knowledge of a particular linguistic property. The experiment is controlled because the rest of the data remains unchanged. When the defined part of speech is present in a sentence, its grammatical form is randomly replaced with the incorrect one with some probability.

While numerous grammatical categories are featured in various languages[4], it can be deduced that each morphosyntactic property may have either a few (e.g., two) or many (particularly, more than four) forms inside it.

In our experiment, we have chosen gender to be the category we "break" as the starting point. It is a widely present feature in grammar. In Russian, it is non-binary, having three values (feminine, masculine and neutral), which seems to be a reasonable trade-off, when tackling the above mentioned problem of language diversity.

Specifically, we focus on adjectives' gender, because it is relatively easy to "spoil" such forms, changing the inflections. While gender is a word-classifying category for nouns in Russian, for adjectives and verbs in the past tense, it is inflectional. Gender is not overtly marked on nouns, but is correlated with the type of declension. Adjectives and verbs agree with nouns by gender, but it is differentiated only in the singular, whereas in the plural, gender is neutralized. Adjective forms of masculine and neuter gender differ only in the nominative and accusative cases.

In our work we largely rely on pymorphy2[5] engine (Korobov, 2015). It's worth mentioning that this morphological analyzer considers pronominal adjectives to be adjectives as well. And since pymorphy2 does not disambiguate analyses, some demonstrative and relative pronouns (e.g. "takoj" 'such', "tot" 'that', "kotoryj" 'which'), participles (e.g. "kuryashchij" 'smoking'), and adverbs (e.g. "ladno" 'fine') were changed too.

[4] https://wals.info/
[5] https://github.com/pymorphy2/pymorphy2

4.2 Probing Data

For probing purposes, there is an option to use a converter from Serikov et al. (2022) which brings the Universal Dependencies data (de Marneffe et al., 2021) to SentEval format (Conneau and Kiela, 2018) so that the files for found morphological categories (e.g. Verb Tense) are created. These files contain train, test and dev tags for sentences as well as particular grammatical labels to predict.

5 Experimental Setup

Pre-trained Neural Language Models: as was mentioned above, we have studied the behavior of the BERT model (Devlin et al., 2019). The idea is to fix random seed and the environment and train two models separately (for the MLM task, in our case) on the practically same dataset, whereas the second version of BERT is fed with "spoiled" data where we randomly changed gender of the long and short forms of the adjectives. Each model was trained for 1kk steps on a single NVidia Tesla GPU V100 32GB with a mini-batch size of 32, which took ~ 35 hours. We also made checkpoints at each 100k steps for the further comparisons.

Pre-Training Data: ~3 GB of conversational texts in Russian, taken from OpenSubtitles (Lison and Tiedemann, 2016), Dirty[6], Pikabu[7], and a Social Media segment of Taiga[8] corpus.

Example of pre-training-data:

В интересные времена живем. ‘We live in interesting times’.

Ведь эти биткойны — первая независимая финансовая система. ‘After all, these bitcoins are the first independent financial system’.

И ее, кажется, даже невозможно уничтожить никак — ни законами, ни киллерами, ни бомбардировщиками, ни подкупами министров. ‘And it seems impossible to even destroy it in any way - neither by laws, nor by killers, nor by bombers, nor by bribing ministers’.

Интересно, какие есть технические возможности. ‘I wonder what technical possibilities there are’.

We have randomly changed ~60% of adjectives’ gender forms in the pre-training data at the step of "spoiling" the dataset for the "broken" model.

Example of "spoiled" pre-training-data:

Ведь эти биткойны — первая независимое финансовый система.

‘After all, these bitcoins are the first(F) independent(N) financial(M) system(F)’[9].

The total distribution of adjectives in the data is shown in Table 1.

[6] https://d3.ru/
[7] https://pikabu.ru/
[8] https://tatianashavrina.github.io/taiga_site/
[9] The letter in brackets indicate the gender form of the word. In a grammatically correct sentence, all three modifiers should agree with the noun and be in the feminine form (F).

Tab. 1. Percentage of full (ADJF) and short adjectives (ADJS) in pre-training data.

(total number of words: 274442625)

POS	Number, gender	Percentage
ADJF	sg, masc	2.86%
	sg, femn	2.86%
	sg, neut	1.93%
	pl	3.09%
ADJS	sg, masc	0.28%
	sg, femn	0.14%
	sg, neut	0.42%
	pl	0.16%
Total		11.76%

Probing Data: UD Russian Taiga[10] converted to SentEval format with UD Parser[11].

Example of probing data:

tr; Fem; Оценка истории крайне интересна !	'The assessment of history is extremely interesting !'
tr; Neut; - Спасибо большое .	'Thanks a lot .'
tr; Masc; - Кто последний ?	'Who is last ?'

Studied Properties: besides probing for adjectives' gender, we have chosen other core parts of speech and their grammatical features. Thus, we studied nouns, verbs and adjectives. Our final list of properties includes: adjectives' gender; nouns' number and case; verbs' aspect, person and tense.

Probe Configuration: all the probing experiments were also conducted with the same fixed seed so that training and test sets always contained identical data and the linear probe's initialization did not mess up the results. In each iteration a linear probing classifier with elastic-net regularization ($\lambda 1$, $\lambda 2 = 0.003$), a categorical cross-entropy loss and AdamW optimiser

[10] https://github.com/UniversalDependencies/UD_Russian-Taiga
[11] https://github.com/AIRI-Institute/Probing_framework/tree/main

(Loshchilov and Hutter, 2017) was used (Durrani et al., 2020). Then the importance of each neuron was calculated according to the learned weights.

Metrics: label prediction accuracy and selectivity.

6 Results

Our analysis splits into two parts: BERT models' comparison and the probing section.

6.1 Pre-Trained BERT Models

The widely used metric for LM's evaluation is perplexity[12]. However, it is devised to estimate the performance of autoregressive models, while BERT is MLM model. Thus, we calculated pseudo-perplexity, following the approach by Salazar et al. (2019). Pseudo-perplexity, in its core, largely relies on the average probability of each token in a given sentence. The interpretation of the metric is as follows: the lower pseudo-perplexity, the better understanding of the language the model has, as it predicts the masked words more accurately, assigning them higher probabilities.

We measured pseudo-perplexity over two subsets. Each of them contained 30000 sentences. They were randomly sampled from RuSentEval data[13,14]. The metrics for "good" and "broken" models on the first dataset are ~41.17 and ~48.11 respectively. The metrics on the second dataset are ~59.73 and ~71.87. The difference between the latter values is higher as, presumably, the second dataset, devoted to estimation of object's gender, contained more adjectives. Nonetheless, the results allow to conclude that the "good" BERT is more successful in language modeling than the "broken" one.

The line plot of loss during training (Fig. 1) shows that models actually learned differently. The loss of the "broken" model is generally a bit higher as the model was confused with the randomly "spoiled" grammatical forms.

[12] https://huggingface.co/docs/transformers/perplexity
[13] https://github.com/RussianNLP/RuSentEval/blob/main/data/sent_len.txt
[14] https://github.com/RussianNLP/RuSentEval/blob/main/data/obj_gender.txt

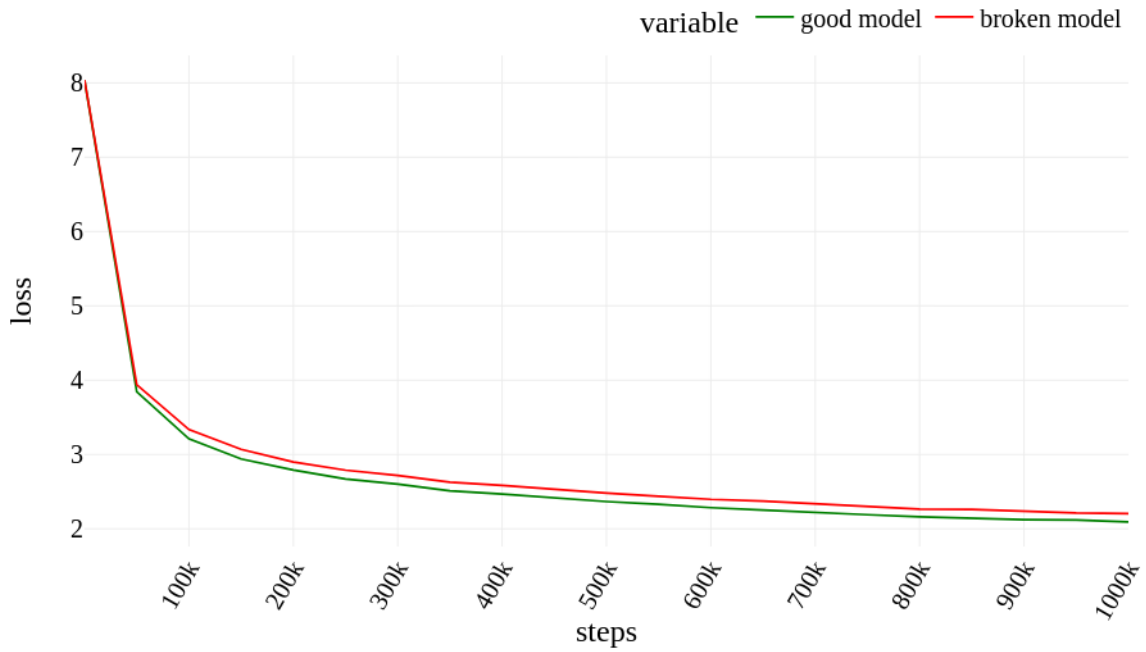


Fig. 1. Loss change during training two BERT models.

Masked language modeling task itself also proves that we have managed to bewilder BERT with conflicting linguistic features. The "broken" model assigns higher probabilities to incorrect grammatical forms when it predicts adjectives (Table 2a).

However, when we mask other parts of speech (Table 2b), the behavior of the models also does not seem to be consistent, as they propose different variants for the same masked word in a sentence, even though it doesn't contain adjectives.

Tab. 2. The way the pre-trained models guess the masked word.

(a) The noun's gender is feminine, so the first model is correct, the second one is not.

Собака очень [MASK]. dog.F-NOM very [MASK][15]		
model	score	prediction
good model	0.284	красивая beautiful-F.NOM
	0.052	добрая kind-F.NOM
	0.048	умная smart-F.NOM

[15] In this table, instead of a translation, a gloss line is given. F – feminine, M – masculine, N – neuter, NOM – nominal case.

broken model	0.028	сильный beautiful-M.NOM
	0.018	красивое beautiful-N.NOM
	0.018	сильная beautiful-F.NOM

(b) The proposed variants and probabilities are different, although the masked word is not an adjective.

Мальчик ходит в [MASK] ежедневно. ‘The boy goes to [MASK] every day’.		
model	score	prediction
good model	0.7486	школу ‘school’
	0.0321	церковь ‘church’
	0.0231	походы ‘hiking’
broken model	0.6992	школу ‘school’
	0.0342	садик ‘kindergarten’
	0.0112	спортзал ‘gym’

6.2 Probing Analysis

We conducted a number of probing experiments based on activations generated by two pre-trained BERT models.

Performance of the Probes

The plot (Fig. 2) indicates that the "good" model is generally better, although the "broken" one performs higher on the noun’s properties.

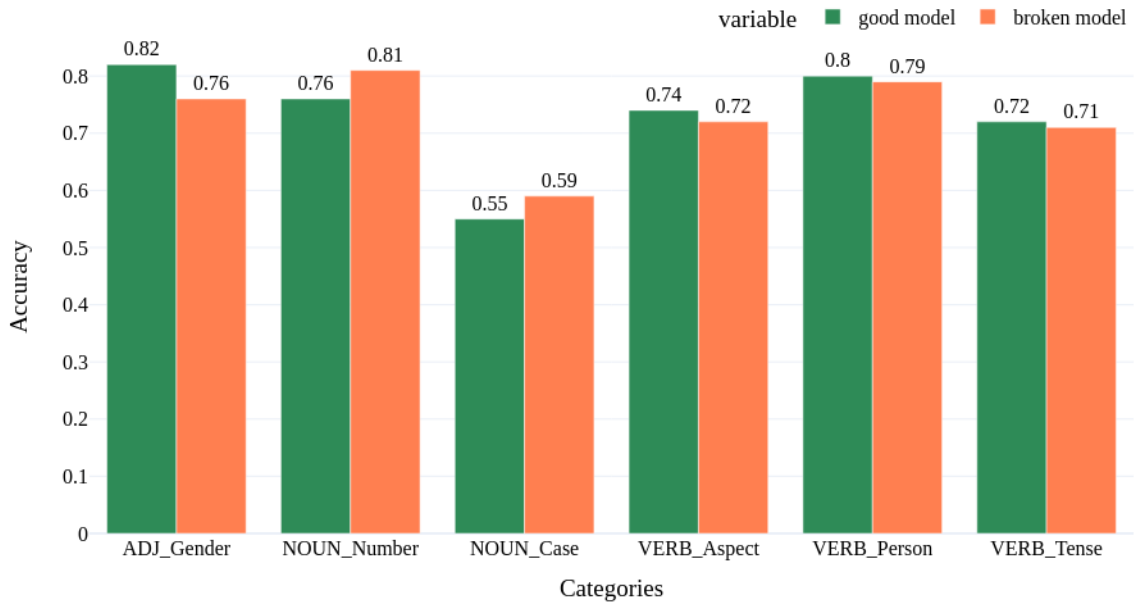


Fig. 2. Test accuracy: models' comparison across the probed grammatical categories

We conducted a control task to make sure the obtained results can be trusted. The difference in performance on "true" and randomly annotated data reveals that the probe does not just memorize the information from the model's representations, but also grasps significant features. Thus, selectivity, which is defined as the target metric subtraction, is rather high across all chosen linguistic categories. Therefore, the probes actually use the encoded knowledge. The comparison is shown in Fig 3.

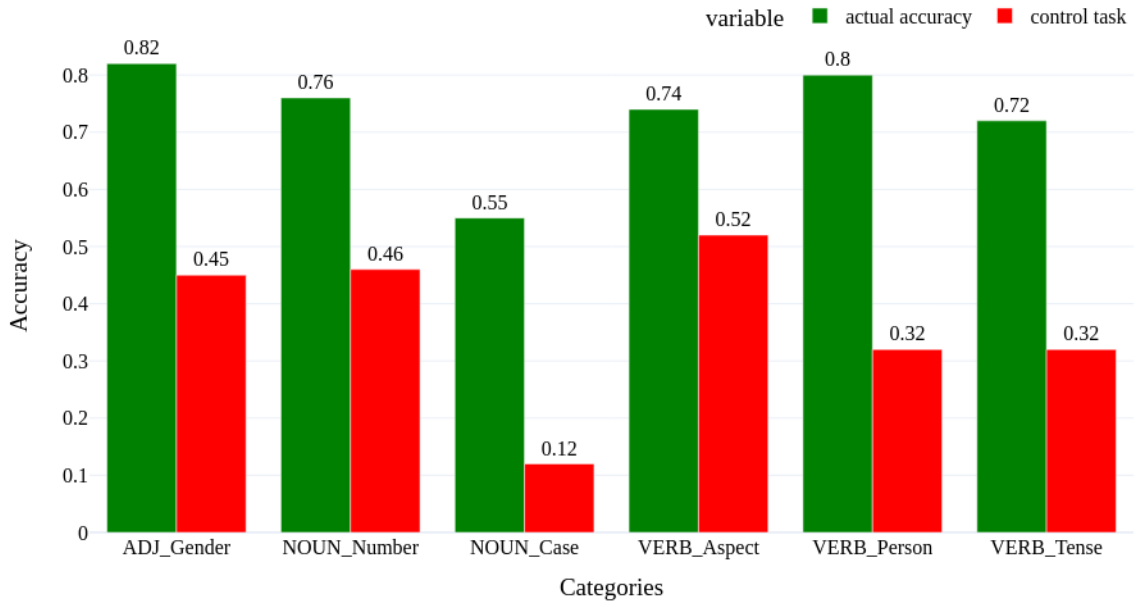


Fig. 3. Actual test accuracy vs control task accuracy for the "good" BERT model.

According to McNemar's Chi-squared test with continuity correction, the difference between the results of the test and control tasks is statistically significant for all categories (p-value < 0.05)

Neuron Sets

While the NeuroX Framework lets analyze particular neurons, we singled out top-20% of neurons by weight mass, bottom-20% and trained probes zeroing-out other neurons. The obtained metrics show (Table 3) that the fewer neurons are considered, the better might be a performance as such an approach probably prevents overfitting. Beyond that, top-neurons indeed contain more information in comparison to bottom-neurons.

Tab. 3. Test accuracy comparison of the performance probe trained on all (total number = 9984) neurons, top-20% by mass and bottom-20% of neurons.

category	all	top-20%	bottom-20%
ADJ Gender	0.82	0.84	0.79
NOUN Number	0.76	0.82	0.81
NOUN Case	0.55	0.6	0.55
VERB Aspect	0.74	0.73	0.71
VERB Person	0.8	0.81	0.63
VERB Tense	0.72	0.73	0.67

Besides, taking a closer look at the number of top-20% of neurons per category (Table 4), we state that the fewer labels the property has, the more weight mass is localized rather than distributed. For example, we see a pattern: 2 labels correspond to ∼ 350-400 neurons, 3 labels – to ∼ 600 neurons and a category with 8 labels requires even a larger number of neurons. Generally, it means that if a linguistic property is complex, the model needs more neurons to encode the knowledge in comparison to a simpler grammatical feature. These results are congruent with previous studies (Durrani et al., 2020).

Tab. 4. Number of classes in each category vs number of top-20% of neurons per category for "good" BERT.

category	classes	neurons
ADJ Gender	3	635

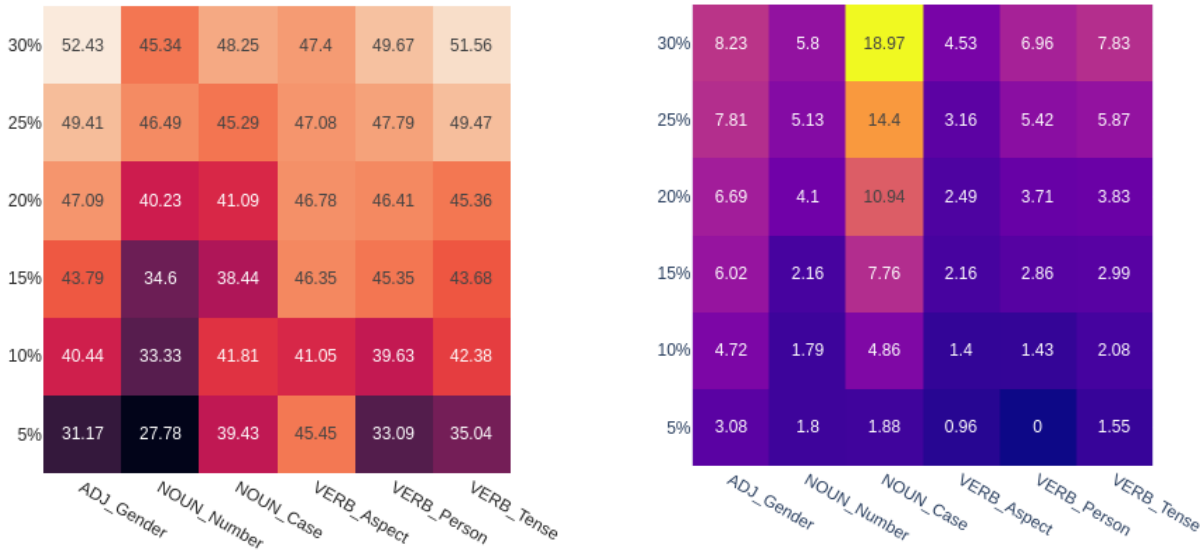
NOUN Number	2	385
NOUN Case	8	1749
VERB Aspect	2	356
VERB Person	3	573
VERB Tense	3	598

Neuron Overlap

At the step of designing the whole experiment, we wanted to detect whether top-N% neurons actually vary across two pre-trained models.

Firstly, we measured an overlap in percentage between two versions of our "good" BERT model. We took the checkpoint at 700k steps and the final one at 1kk steps. The overlap displays that the model maintains continuity, as the intersection is significant (Fig. 4a).

Secondly, we calculated the same overlap across "good" and "broken" models. According to the heatmap (Fig. 4b), it is not large, which raises the question whether the models are actually that different or if the comparison should be made in another way.



(a) (b)

Fig. 4. Percentage of top-N% neurons (by weight mass) overlap: a) comparison between "good" BERTS after 700k and 1kk training steps; b) comparison between “good” and "broken" BERTS after 1kk steps.

Layers: Distribution and Evaluation

The layer-wise distribution of the salient neurons (Fig. 5) provides more information about the way grammar is encoded. Our assumption is that the number of significant neurons per layer can be helpful when we compare "sibling" models. Although we can't say that distributions are nearly identical, the trends of both "good" and "broken" models for most of the categories generally do not contradict each other.

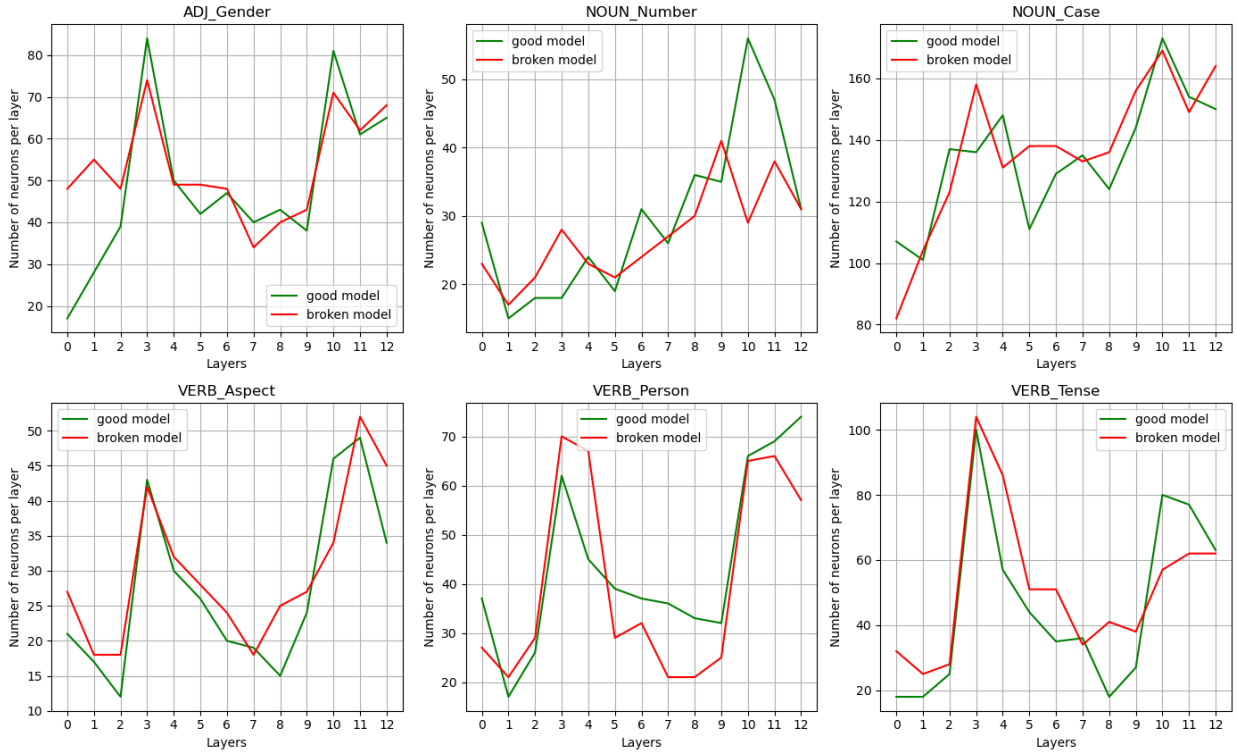


Fig. 5. Top-20% of neurons per category: layerwise distribution for two probed BERT models after 1kk steps.

Training the probing classifiers, we expected a large drop in performance on our major target category – adjective's gender. Although the probe based on the "good" model actually makes better predictions as shown in the layer-wise accuracy plot (Fig. 6), the "broken" model also does well. The possible explanation is that the probe has found patterns in the sentence representations of the correspondent to the target adjective noun, which is the head of the noun phrase. Still, we see a peak in performance at layer three for both models. The result is in line with Fig. 5, as the third layer has more top neurons than any other one for both "good" and "broken" BERTs.

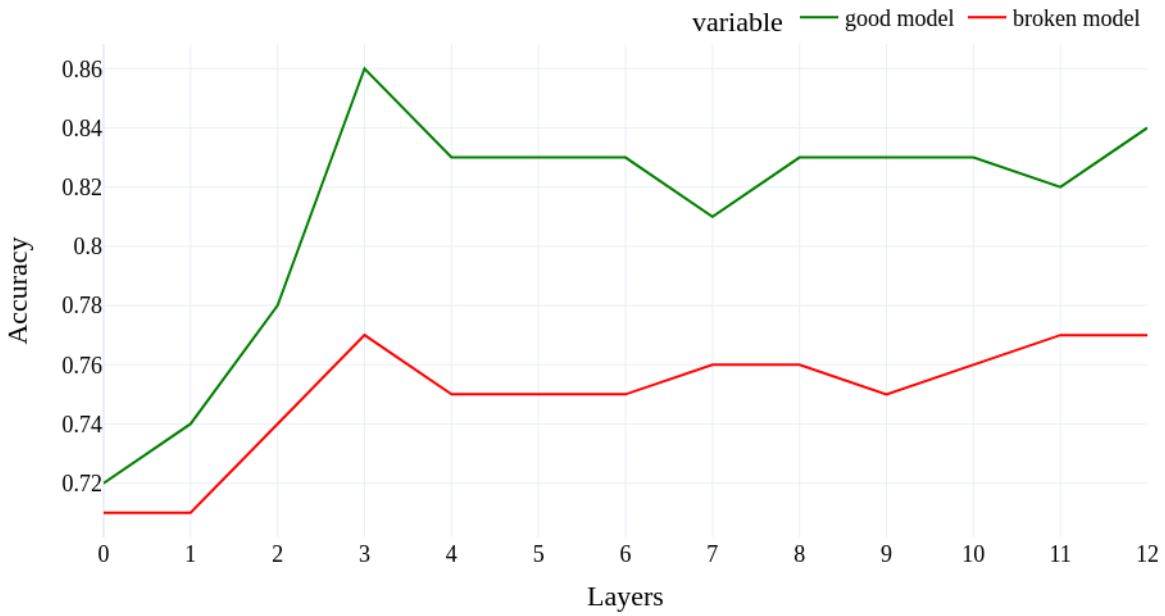


Fig. 6. Layerwise test accuracy: models' comparison for adjectives' gender category.

7 Future Work

The work can be continued so that one can experiment with other architectures, languages, linguistic categories and probing methods.

Our plan is to conduct a similar experiment where values of a grammatical category are always changed into the same form rather than randomly. Thus, we will be able to "turn off" the grammatical category in the "broken" model: it will not know that the particular part of speech can inflect for it.

Speaking about the probing part, in our study we focus on the sentence representations. We want to do conduct future probing experiments on word-annotated text-label pairs.

8 Conclusions

In this work we propose a new approach to the probing task, which aims to extract features relevant to a specific linguistic category from a deep neural network's representations through a novel experimental setup. Our findings are stated as follows:

- it is a promising technique to train two architectures in the same fixed conditions, but on slightly different data in terms of grammatical correctness; it allows to draw a comparison between activations generated by the models;

- a subset of top-N% neurons is much more informative than all number of neurons for a linguistic property as long as it partly solves the overfitting problem during probing procedure;
- such a subset is smaller for "simple" categories, while its size, probably, grows linearly as the number of labels increases;
- "complex" properties are distributed because it is harder for the model to internalize them than some "easy" ones, which, interestingly, is understandable from the perspective of "human" interpretation (though, our conclusion confirms previous studies in this field);
- BERT model is rather coherent during pre-training, so that it is enforced to encode grammar in approximately the same subset of neurons at each iteration;
- the layerwise top-N% neuron distribution is useful for comparison analysis between two pre-trained "sibling" models.

Limitations

Firstly, in our study we have dealt primarily with adjectives' gender category and while a full adjective is a part of a noun phrase, it is a dependent, not a head of NP. It is strongly advised to conduct further experiments "spoiling" a head of a phrase (or a head with its dependents) in data in order to get the most solid results.

Secondly, our approach requires a morphological analyzer which can parse words and change their grammatical forms. Though, a solution for low-resource agglutinative languages is the usage of transducers. But the better the analyzer disambiguates homonyms, the more accurate conclusions about the behavior of the "broken" model can be made.

Thirdly, scaling the approach to the architectures with a vast number of parameters is bound to require lots of computational power and, thus, be costly.

References

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. Naturalistic Causal Probing for Morpho-Syntax. Transactions of the Association for Computational Linguistics, 11:384– 403.

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

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019a. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):6309–6317.

Fahim Dalvi, Avery Nortonsmith, Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019b. Neurox: A toolkit for analyzing individual neurons in neural networks. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):9851–9852.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. Computational Linguistics, 47(2):255–308.

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.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4865–4880, Online. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. Transactions of the Association for Computational Linguistics, 9:160–175.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. 2021. Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids' representations. In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 375–388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1828–1843, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. CoRR, abs/1909.03368.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, Analysis of Images, Social Networks and Texts, volume 542 of Communications in Computer and Information Science, pages 320–332. Springer International Publishing.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 923– 929, Portorož, Slovenia. European Language Resources Association (ELRA).

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. CoRR, abs/1711.05101.

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.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2019. Pseudolikelihood reranking with masked language models. CoRR, abs/1910.14659.

Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova, and Tatiana Shavrina. 2022. Universal and independent: Multilingual probing framework for exhaustive model interpretation and evaluation. In Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 441–456, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Ekaterina Voloshina, Oleg Serikov, and Tatiana Shavrina. 2022. Is neural language acquisition similar to natural? a chronological probing study. In Computational Linguistics and Intellectual Technologies. RSUH.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1651– 1661, Florence, Italy. Association for Computational Linguistics.

Ksenia E. Chistyakova
HSE University (Moscow, Russia).
E-mail: ksevgch@gmail.com

Tatiana B. Kazakova
HSE University (Moscow, Russia). Laboratory for Arctic Social Sciences and Humanities. Research Assistant
E-mail: tanusha.kazakova@gmail.com

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.