Behavior of Modern Pre-trained Language Models Using the Example of Probing Tasks

Ekaterina Kalyaeva
kalyaevarina@mail.ru

Oleg Durandin
oleg.durandin@gmail.com

Alexey Malafeev
amalafeev@yandex.ru

National Research University Higher School of Economics
Nizhny Novgorod, Russia

Abstract

Modern transformer-based language models are revolutionizing NLP. However, existing studies into language modelling with BERT have been mostly limited to English-language material and do not pay enough attention to the implicit knowledge of language, such as semantic roles, presupposition and negations, that can be acquired by the model during training. Thus, the aim of this study is to examine behavior of the model BERT in the task of masked language modelling and to provide linguistic interpretation to the unexpected effects and errors produced by the model. For this purpose, we used a new Russian-language dataset based on educational texts for learners of Russian and annotated with the help of the National Corpus of the Russian language. In terms of quality metrics (the proportion of words, semantically related to the target word), the multilingual BERT is recognized as the best model. Generally, each model has distinct strengths in relation to a certain linguistic phenomenon. These observations have meaningful implications for research into applied linguistics and pedagogy, contribute to dialogue system development, automatic exercise making, text generation and potentially could improve the quality of existing linguistic technologies.

1 Introduction

As is well-known, 2018 saw a breakthrough in natural language processing (NLP) with the advent of several novel pre-trained language models, including BERT (Devlin et al, 2018). These models are capable of fine-tuning, that is, additional training on a smaller dataset for a specific task. Such models develop a certain degree of understanding natural language and therefore require further studies (Zhang et al, 2019; Wallat et al, 2020). To this end, the so-called probing tasks are commonly used, such as testing the model's ability to identify semantically coherent sentences in running text. A considerable amount of recent research (Gulordava et al, 2018; Salazar et al, 2019, Sun et al, 2019) was devoted to the general language modelling problem and also to masked language modelling. Other scholars (Devlin, 2018; Gong, 2019; Rogers et al, 2020) explored BERT specifically. However, there is a lack of studies concerning BERT's behavior on Russian language material. In addition, existing papers discuss the problem without taking into consideration systematic linguistic features.

This paper aims to explore the behavior of the pre-trained BERT language model using the example of the diagnostic task of masked language modeling for the Russian language and give a linguistic interpretation of the cases when the model shows unsatisfactory results. This study will focus on new language data with experiment design based on linguistic theories[1]. The problem will be discussed in terms of the language modelling concept, cognitive science, theory of language, in particular the language frames theory, semantic roles concept, and also negations processing depending on a context.

2 Related Work

Bengio and co-authors (Bengio et al, 2003) suggest the following definition of a language

[1] All datasets and the Jupyter notebook with the experiments are made available as a GitLab repository: https://gitlab.com/lieutkat/linguistic-experiments-with-bert/

Proceedings of Recent Advances in Natural Language Processing, pages 664–670
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_075

model: probability distribution over word sequences. Such models can be the basis for solving a large number of NLP tasks, with the help of fine-tuning the general language models. The pre-trained language model weights already encode a lot of information about natural language. Specifically, a pre-trained model provides features of semantics, syntax, and “verbal knowledge” which may be transferred from general to specific tasks (Roberts et al, 2020; Manning et al, 2020). In the past several years, the BERT model and its modifications have become widely used in the natural language processing community (Liu et al, 2019; Lan et al, 2019). The model can be applied to all general types of tasks: single sentence prediction, sentence pair classification, question answering and sentence tagging. As BERT is inherently an encoder for language information, it often plays the role of a text feature extractor (Rogers et al, 2021).

Some authors (Conneau et al, 2018; Kim et al, 2019) focus on different types of tasks regarding language models, namely probing tasks. These authors suggest options for surface information (e.g., recovering a word from its embedding), syntactic information (e.g., the model sensitivity to word order change), semantic information (e.g., identification of main-clause verb tense). Their basic application is testing the pre-training effect: how pre-trained models encode various language phenomena. Currently, there are lists of linguistic capabilities available that allow to explore the of behavior of language models (Ribeiro et al, 2020). Thereby, the main benefit of probing tasks is the analysis of linguistic knowledge that can be extracted from sentence representations (Hewitt et al, 2019).

Other authors (Devlin et al, 2018; Song et al, 2019) suggested a new probing task for language models that is nowadays known as masked language modelling. The key difference is that the model masks some random words from a sequence and then predicts them again based on the contextual information, both left and right. Ultimately, it was suggested to train with masked language modeling, and then fine-tune for specific tasks (Wu et al, 2019; Salazar, 2019).

Differences in human and machine understanding of language were investigated in another paper (Ettinger, 2020), which inspired our experimental setup. The author uses psycholinguistic tests to find out the sensitivity of the model to such phenomena as hypernymy, semantic roles and negation. The tests revealed weaknesses in the language model and proved that computational models have great potential for natural language understanding.

As was pointed out in the previous section, despite this scholarly interest for BERT, there is a lack of studies with a strong linguistic basis. Additionally, previous studies have tended to focus primarily on English language material. The present study is based on a Russian-language dataset and uses theories from cognitive linguistics and the theory of language.

3 Linguistic Basis

From the linguistic perspective, the work of language models is based on the concept of semantic roles (Fillmore, 1976). Speaking about the subject of action, it can have both an agentive (or active) position, and a non-agentive (or inactive) position (Uskova, 2012). Ch. Fillmore also developed the theory of presupposition, which is important for this study. It is the preliminary knowledge that is responsible for the semantic correctness of the utterance. For example, the sentence Snow is expected in Moscow on February 30th, is incorrect due to the fact that it includes a false presupposition on February 30th.

Meanwhile V.Z. Demyankov identifies several types of presuppositions. In our experiment, we will use pragmatic and logical types. Pragmatic presupposition is the conditions and contexts that must be present in order for the speaker's intention to be realized correctly. The semantic presupposition characterizes the relationship between a sentence and the proposition it expresses.

Another linguistic aspect that requires our consideration is negation. It is considered a semantic primitive integrated into the grammatical and lexical systems of all languages of the world (Paducheva, 2011). We will explore predicate negation.

Speaking about studies of language models, it is necessary to mention the type of lexical paradigmatic relations, such as hypo-hyperonymic. This type of relations reflects the direction of human thinking to systematize lexical units and non-linguistic structures behind them and bring them to a hierarchical form (Kuznetsova, 1989). This fact can be used to evaluate the performance of the language model in such a way that the resulting metric value is interpretable.

So, after considering some linguistic theories and concepts, we came to the term “behavior of the language model". According to the Philosophical Encyclopedia (Ilyichev et al, 1983), behavior is a way of reacting to any influence. In this paper, the behavior of a language model is understood as text data that the model provides as output under certain conditions of its use, namely, when testing or applying to a problem.

4 Datasets and Methods

The final dataset consists of three parts, each containing 50 sentences, which were collected using a linguistic observation method. Each part was designed to test and evaluate a certain aspect of the functioning of the language model, namely: common sense inference, the interpretation of semantic roles, and processing negatives depending on the context. Table 1 shows a sample of the data for semantic role interpretation.

	content	target
1	В процессе разговора я вдруг заметил, что она частенько [MASK] переходит на украинский язык. During our conversation, I suddenly noticed that she often [MASK] to Ukranian.	переходит switched
2	В процессе разговора я вдруг заметил, что украинский язык частенько [MASK] в ее речи. During our conversation, I suddenly noticed that the Ukranian language was often [MASK] by her.	слышится used
3	Она [MASK] местным украинским землячеством. She [MASK] the local Ukranian community.	руководила leads
4	Местное землячество [MASK] ею. The local Ukranian community was [MASK] by her.	руководилось led

Table 1. Example of a subcorpus for checking the interpretation of semantic roles.

There were two main resources for collecting data. Firstly, educational texts for teaching Russian as a foreign language for B1+ learners, with the help of which the "content" column was filled. Secondly, the National Corpus of the Russian language (RNC)[2], in which the expected words for contexts were selected through a semantic search. For instance, the selection of contexts for nouns with time semantics was carried out by entering the following characteristics: *r:abstr & (t:time:period | t:time:moment).*

In the case of semantic roles, complex or simple common sentences were selected from educational texts, the predicate was replaced with a token mask, and the linguistic unit, which had the role of an agent, took on the role of a patient (or experiencer).

To check common sense inference, we selected two consecutive sentences from the educational texts where one word in the second sentence was manually replaced with a mask. Then, using semantic search in the RNC, we selected a word of a related topic and a word of a more general topic with respect to the expected topic. Next, the subcorpus with negations was compiled as follows: in the RNC, using semantic search, we selected sentences with an adjective at the end, and masked the adjective. Finally, the sentence was copied and transformed into a negative one, and the adjective was replaced by its antonym.

The dataset was examined using a Jupyter notebook in the Python programming language with the help of the libraries tensorflow, pytorch and transformers. To perform a comparison, the BERT DrMatters [3] and BERT DeepPavlov [4] models for the Russian language, and the multilingual BERT[5] were selected.

- The BERT model (bert-base-multilingual-cased) includes 12 layers, 768 hidden layers, 12 heads of the attention mechanism and 179 million parameters. Trained on Wikipedia texts for 104 languages.
- The RuBERT (or BERT DeepPavlov) model was created by the team of the Moscow Institute of Physics and Technology. It contains 12 layers, 768 hidden layers, 12 heads of the attention mechanism and 180 million parameters. The model was trained on the Russian-language Wikipedia and news.
- BERT DrMatters model is based on the BERT DeepPavlov. There is a lack of information about all the other characteristics.

For the assessment, the RuWordNet model was loaded using the ruwordnet package. We received lists of hyperonyms and hyponyms for the target

[2] To access the RNC search interface see https://ruscorpora.ru/old/en/search-main.html

[3] See https://huggingface.co/DrMatters/rubert_cased

[4] See http://docs.deeppavlov.ai/en/master/features/models/bert.html

[5] See https://huggingface.co/bert-base-multilingual-cased

word, then checked which of the prediction words are included in this list and counted them. Subsequently, the resulting list of numbers was normalized for convenient processing. The obtained data can be interpreted as the proportion of words in the prediction that are semantically related to the target word (that is, there are hypo-hyperonymic relations between them). Using such a metric, it is possible to understand whether the model is trying to put the correct group of objects in the place of the mask, that is, to test the model's ability to generalize and differentiate.

5 Results

5.1 Quality of Predictions

Table 2 provides information about the quality of the models' predictions for several types of experiment, the measure is normalized RuWordNet-based. The largest values for each type of experiment are highlighted in gray.

aspect	BERT Multilingual	BERT DeepPavlov & DrMatters
common sense	0.1	0.1
semantic roles	0.12	0.1
negations	0.12	0.1
negations (aff)	0.17	0.15
negations (neg)	0.17	0.15

Table 2. Measures based on RuWordNet for multilingual and Russian models (aff – affirmative context, neg – negative context).

The advantage of the measure based on RuWordNet is the fact that a successful prediction is not only the target word itself, but also many of its hyponyms and hyperonyms. For this reason, the values tend to be quite interpretable. Overall, the multilingual model in experiments with semantic roles and negations is slightly ahead of both Russian ones, which have equal values for each aspect. Moreover, no differences were found between the processing of the two types of contexts in the negation experiment. The results of the linguistic analysis are presented below.

5.2 Linguistic Analysis: Common Sense Inference

In case of **multilingual BERT**, the influence of the pragmatic presupposition is strong, while the role of the logical presupposition is minimal (see Table 3, example 1). It can be assumed that the model uses the acquired background knowledge about historical facts. At the same time the demonstration of background knowledge dominates the observance of grammatical correctness of the prediction (see Table 3, example 2).

	sentence	target	model's predictions
1	Свое имя медуза получила из-за сходства с шевелящимися волосами-змеями легендарной Медузы Горгоны из [MASK] Греции. The medusa got its name because of its similarity to the moving hair-snakes of the legendary Gorgon Medusa from [MASK] Greece.	мифологии mythology	['столицы', 'города', 'из', 'Новой', 'Великой', 'народов', 'Западной', 'театра', 'жителей', 'династии'] ['capital', 'city', 'from', 'New', 'Great', 'peoples', 'Western', 'theater', 'inhabitants' , 'dynasty']
2	Рост Наполеона был выше среднеевропейского. Историки давно закрепили за французским [MASK] прозвище Маленький Капрал. Napoleon's height was higher than the average European. Historians have long fixed the nickname Little Corporal for the French [MASK].	полководцем commander	['Наполеон', 'человеком', '-', '.', 'орденом', 'как', 'на', 'именем', 'в', 'это'] ['Napoleon', 'man', '-', '.', 'order', 'as', 'on', 'name', 'in', 'this']

Table 3. Examples from qualitative analysis, multilingual BERT (common sense).

Both models for the Russian language have problems with the pragmatic presupposition: sometimes it is insufficient to fulfill the predictions. Inaccuracies in the predictions occur under the influence of the logical presupposition when it is more pronounced than the pragmatic one (see Table 4). In the example below, the models incorrectly interpret the logical presupposition and perform a pronominal replacement of the subject ("the boy").

	sentence	target	model's predictions
1	Парнишка достал чехол, вытащил ружье и начал его собирать. Потом полез за [MASK],	патронами cartridges	['ним', 'собой', 'ними', 'патроны', 'ружье',

	рассыпал их, начал торопливо выбирать нужные. The boy took out a case, pulled out a gun and began to assemble it. Then he reached for the [MASK], scattered them, and began hurriedly choosing the right ones.		'руки', 'патронами', 'столом', 'доской', 'рамки'] ['him', 'himself', 'them', 'cartridges', 'gun', 'hands', 'cartridges', 'table', 'board', 'frames']

Table 4. Examples from qualitative analysis, BERT DeepPavlov and BERT DrMatters (common sense).

5.3 Linguistic Analysis: Interpretation of Semantic Roles

In the predictions of **the multilingual BERT mode**l, the target words themselves are almost absent, but there are more common synonyms for them (see Table 5, example 1). For the case of the non-agentive position of the subject, the model is practically unable to predict the target word correctly (see Table 2, example 2).

	sentence	**target**	**model's predictions**
1	В процессе разговора я вдруг заметил, что она частенько [MASK] на украинский язык. During our conversation, I suddenly noticed that she often [MASK] to Ukranian.	переходит switched	['говорит', 'говори', 'вышла', 'пишет', '-', 'написана', 'носит', 'говорить', 'выходит', 'писал'] ['speaks', 'speak', 'came out', 'writes',' -', 'written', 'wears', 'speak', 'comes out', 'wrote']
2	Местное украинское землячество [MASK] ею. The local Ukranian community was [MASK] by her.	руководилось led	['с', 'над', 'перед', 'за', 'под', '-', 'между', 'было', 'в', 'к'] ['with', 'over', 'before', 'for', 'under',' -', 'between', 'was', 'in', 'to']

Table 5. Examples from qualitative analysis, multilingual BERT (semantic roles).

As for **BERT DeepPavlov and BERT DrMatters**, when the subject is in the agentive position, the predictions often include the target word (see Table 6, example 1). Moreover, in the situation of the non-agentive position of the subject, the short length of the sentence negatively affects the predictions (see Table 6, example 2).

	sentence	**target**	**model's predictions**
1	Стресс можно [MASK] с помощью медуз. Stress can be [MASK] with the help of jellyfish.	снять reduced	['вызвать', 'снять', 'преодолеть', 'предотвратить', 'вызывать', 'пережить', 'снизить', 'переносить', 'уменьшить', 'облегчить'] ['cause', 'remove', 'overcome', 'prevent', 'cause', 'survive',' reduce', 'transfer', 'reduce', 'facilitate']
2	Завтра вечером в парке я буду [MASK] другом. Tomorrow evening I will be [MASK] by a friend in the park.	встречен met	['с', 'моим', 'в', 'совсем', 'быть', 'заниматься', 'хорошим', 'на', 'работать', 'спать'] ['with', 'my', 'in', 'at all', 'be', 'engage', 'good', 'on', 'work', 'sleep']

Table 6. Examples from qualitative analysis, BERT DeepPavlov and BERT DrMatters (semantic roles).

5.4 Linguistic Analysis: Negations Depending on the Context Type

Multilingual BERT's predictions contain a lot of noise in the form of subwords, UNK-tokens, and punctuation marks. In the predictions without noise, there are often explicit or implicit negation components (e.g., “no”, “absent”, “lack”). Thus, we can suppose that at a high level of abstraction, the model has acquired information about the nature of the context, but it does not have enough representations to express it (see Table 7).

	sentence	**target**	**model's predictions**
1	При отсутствии грунтовки стена не будет одинаково связывать влагу в различных местах, даже если поверхность [MASK]. In the absence of a primer, the wall will not bind moisture	гладкая smooth	['нет', 'отсутствует', 'должна', 'поверхности', 'имеется', 'существует', '-', 'т', 'установлена', '##ная'] ['no', 'missing', 'should', 'surfaces', 'exists', 'exists',' -',

	equally in different places, even if the surface is [MASK].		't', 'installed', '# # naya']

Table 7. Examples from qualitative analysis, multilingual BERT (negations).

Many predictions of **Russian language BERT models** are similar for both affirmative and negative contexts. In the results for the affirmative context, the domination of some positive semantics is observed, despite the fact that the target word is not predicted (see Table 8).

	sentence	**target**	**model's predictions**
1	Газеты говорят, что город выстоит, что он [MASK]. The newspapers say that the city will stand, that it [MASK].	сильный strong	['готов', 'победит', 'свободен', 'есть', 'победил', 'будет', 'выиграет', 'разрушен', 'опасен', 'хочет'] ['ready', 'will win', 'free', 'is', 'won', 'will', 'will win', 'destroyed', 'dangerous', 'wants']

Table 8. Examples from qualitative analysis, BERT DeepPavlov and BERT DrMatters (negations).

6 Discussion

Crucially, the models are unable to understand most of the background information (mostly pragmatic) out of context. Additionaly, the models actually handle the subject better in a more typical, i.e., agentive, position. There are mistakes in processing negatives: all models often choose options for negative contexts that are relevant for affirmative contexts.

Comparing the results of the linguistic analysis of predictions, it is impossible to conclude unambiguously which of the models works better with Russian-language material. Generally speaking, the multilingual model performs better in working with background knowledge and implicit information hidden in the models and for tasks in which strong restrictions are not imposed on the predicted words. The Russian-language model is more suitable for modeling logical relationships and analyzing the subject in its typical position. However, both models do not seem suitable for the problem of negation processing since they consider only one kind of context.

7 Conclusion and Future Work

This paper has investigated the pre-trained BERT language models with probing tasks. As far as we are aware, this is the first time that BERT was explored by professional linguists based on Russian-language material. As a result of this research, we have listed the frequent mistakes made by the BERT model in the masked language modelling task and conducted their analysis. In contrast to previous studies (Ettinger, 2020), where the focus was on psycholinguistic diagnostics and comparison between machine and human performance, we made our study more linguistically grounded by adding knowledge from theory of language. While conducting the research, we created a unique dataset that represents several cognitive phenomena with special annotations. The dataset is made available to the community. In the future, we plan to expand the dataset and annotations to cover more cognitive and linguistic aspects.

Acknowledgments

This work is supported by RSF (Russian Science Foundation) grant 20-71-10010.

References

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research, 3,* 1137-1155.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070.*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Ettinger, A. (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.

Fillmore, C. J. (1976, October). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech* (Vol. 280, No. 1, pp. 20-32).

Hewitt, J., & Manning, C. D. (2019, June). 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) (pp. 4129-4138).

Kim, N., Patel, R., Poliak, A., Wang, A., Xia, P., McCoy, R. T., ... & Pavlick, E. (2019). Probing what different NLP tasks teach machines about function word comprehension. *arXiv preprint arXiv:1904.11544*.

Kuznetsova, E. V. (1989). Lexicology of the Russian language. High school.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences, 117(48*), 30046-30054.

Prokhorov, Aleksandr Mikhaĭlovich. "Great Soviet Encyclopedia." (1970).

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118.*

Roberts, A., Raffel, C., & Shazeer, N. (2020). How Much Knowledge Can You Pack Into the Parameters of a Language Model?. *arXiv preprint arXiv:2002.08910.*

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics, 8,* 842-866.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2019). Masked language model scoring. *arXiv preprint arXiv:1910.14659.*

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal, 27(3)*, 379-423.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450.*

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950.*

Wallat, J., Singh, J., & Anand, A. (2020). BERTnesia: Investigating the capture and forgetting of knowledge in BERT. *arXiv preprint arXiv:2010.09313.*

Wu, X., Zhang, T., Zang, L., Han, J., & Hu, S. (2019). "Mask and Infill": Applying Masked Language Model to Sentiment Transfer. *arXiv preprint arXiv:1908.08039.*

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129.*