NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

As a manuscript

Oleg Serikov

INTERPRETING KNOWLEDGE REPRESENTATION IN NLP MODELS

Dissertation Summary

for the purpose of obtaining academic degree

Doctor of Philosophy in Philology and Linguistics

Academic supervisor:
Candidate of Sciences
Anastasia Bonch-Osmolovskaya

Moscow — 2024

Introduction

1.1 Topic, Content, and Structure of the Work

Large language models are the flagship approach in solving natural language processing and artificial intelligence tasks. Such models are equipped with mechanisms that map contexts (sequences of words or tokens) into a vector space. Context vectors serve as informative features when solving various machine learning tasks. Systems built in this manner demonstrate high performance in specific tasks[1] [1; 2] and achieve results comparable to human performance in general artificial intelligence tasks [3; 4]. A significant number of applied products are also based on large language models [5; 6].

One of the key factors that make large language models unique is their ability to be fine-tuned based on large data collections. Due to the large amount of training data used (training corpora for language models can reach hundreds of gigabytes), language models are capable of accurately processing complex linguistic patterns and generating naturally sounding texts. Pre-training on massive text corpora is sufficient for these models to solve a wide range of downstream tasks with minimal fine-tuning [7; 8]. Thus, pre-trained language models are used to tackle a broad spectrum of tasks such as sentiment analysis, named entity recognition, text classification, and machine translation. There is a growing popularity of language models offered as standalone services (similar to systems like Instruct GPT [9], GPT4 [10], and GigaChat [11]). The use of large language models provides an increase in quality even in non-textual data processing algorithms (video, audio, tabular data) for solving multimodal tasks [12; 13].

The widespread application of language models sparks a particular interest in their predictability and justification of their behavior. Explainability of these systems is crucial for assessing the risks associated with their application. Additionally, the ability to systematize patterns in model behavior is potentially useful for their enhancement: language model training requires extensive resources and thus relies on multifaceted preliminary analysis. Systems used repeatedly in language

[1] https://www.dialog-21.ru/evaluation/

processing tasks rely on previously assimilated knowledge in the form of billions of numerical variables. This format of information representation complicates the explanation and prediction of their behavior. Consequently, the interpretation of knowledge representation in neural network models for natural language processing becomes relevant. The term "knowledge" here and onwards is understood in the same sense as used in knowledge graph methodologies, stemming from relational semantic structures: a pair <terms, a set of semantic relations between them>, reflecting an extra-linguistic reality of an ontological or encyclopedic nature. In the experiments conducted within the scope of the dissertation research, knowledge will be represented as a set of structures in the form of <relationship/predicate type, its participants>, or similar ones. The resource Wikidata has been chosen as the source of such data.

Research towards explaining the behavior of constructed language models is conducted in directions involving assessing and interpreting model behavior in specific tasks, as well as it has been conducted towards a general interpretation of such systems using probing techniques (investigating the relationship between the behavior of individual model elements and some property of processed data). The majority of probing studies are concentrated on analyzing the models' acquisition of language grammar, while semantic probing is less frequent.

The **Objective** of this work is to interpret linguistic and extra-linguistic knowledge accumulated in language models.

To achieve the set goal, it became necessary to solve the following **Tasks**:

1. Analysis and systematization of relevant research directions, methods, and tools for interpreting language models, their key findings, and weaknesses.
2. Determination of data requirements necessary for conducting interpretation research on knowledge in language models.
3. Development of a methodology for conducting interpretation research on knowledge in language models based on the conducted domain review and established data requirements.
4. Verification of the applicability of the layered language model to analyze language representation in neural network models.
5. Conducting semantic and syntactic interpretation of language models.
6. Conducting extra-linguistic interpretation of language models.

7. Publication of interpretation tools and knowledge research in language models in open access, implementing the embodiment of the methodology.

Key Points for Defense:

1. Among the existing techniques for structural interpretation of language models, the predominant majority of works are dedicated to interpreting grammatical structures that arise in the models. The embodiment of semantic structures and structures describing knowledge in language models remains underexplored.
2. To conduct research addressing this gap, annotating hidden structures in texts proves useful. Sources for such annotation can include both linguistic resources (e.g., sketches annotated with semantic roles) and more general resources (knowledge graphs correlated with texts).
3. The behavior of various segments of the neural network is correlated with the presence of different types of predicates encoding knowledge. For semantically similar predicates, the neural network segments show the highest correlations.
4. To ensure comparability of interpretation results across different language models, often belonging to different languages, a toolkit standardizing the conduct of such research is necessary.

Scientific Novelty:

1. For the first time, approaches based on semantic roles and knowledge graphs are applied to interpret language models.
2. A methodology for explicitly analyzing attention heads specialization in language models is proposed for the first time.
3. The hierarchy of predicate types populating the knowledge graph was determined for the first time by interpreting the language model in a deterministic manner.

The **Scientific Significance** of the research is driven by the development of a methodology applying a rich theoretical linguistics toolkit to interpret semantic information and knowledge extracted from language models. The contribution also lies in devising an approach for a deterministic method to construct hierarchies of relationship types present in knowledge graphs.

Practical Significance of the research is ensured by the publication and repeated use in research work of an open toolkit that allows for unified

conduct of multiple interpretation studies on language models. Specifically, the interpretation study of the BLOOM model [14], conducted using this toolkit, enabled the identification of strengths and weaknesses of the multilingual model and characterized the model's multilingualism from a linguistic theory standpoint.

The **Reliability Level** of the obtained results is ensured by successful publication in journals and peer-reviewed conference proceedings. The findings align with results obtained by other researchers, contributing to the evolving scientific discourse on interpreting neural network language models.

Work Validation is ensured by the successful application of the obtained results by the scientific community in subsequent research [14].

1.2 Interpretation of Language Models

In recent years, research aimed at interpreting language models has gained significant popularity. In 2022, at one of the largest conferences dedicated to natural language processing, EMNLP 2022 [15], the field of interpreting "black-box" models became the largest in terms of publication volume [16].

Works such as [17] and [18] provide reviews of the area of interpreting language models from the perspective of reflected language structures. These reviews encompass a broad spectrum of interpretation studies focusing on grammar and language typology diversity. However, semantic resources and knowledge graphs are applied much less frequently in these studies. Addressing this issue are the research and developments described in works [18], [17], [19], [20], [18], [21], [22], [14], presented respectively in the appendices to this work.

1.2.1 Review of Approaches to the Problem of Interpreting Language Models

Many remarkable abilities of large language models stem from the neural network having already identified the regularities needed to solve practical tasks

before it is applied to the task itself [23]. In classical machine learning, tasks are solved in a different order: dependencies are first annotated in the data, then the model is trained on these, and only after that it becomes suitable for solving the target task. One way to explain the premature presence of useful abstractions and systemic characteristics in large neural network language models is the concept of self-organization. This concept suggests that complex systems tend to develop ordered patterns through the interactions of their components [24]. The concept of self-organization is used to systemically describe the workings of the human brain [25] and the behavior of biological species living in societies, such as ants [26].

A significant area of current research is devoted to interpreting language models from a linguistic perspective. The motivation behind such studies is to understand to what extent models "understand" language, or more precisely, to what level of generalization the model's language representations coincide with the generalizations and theoretical constructs accepted in theoretical linguistics. It is important to note that both the conceptualization of language performed by the model and the conceptualizations accepted by linguists are constructs and do not have material analogs in the world. Analyzing the regularities governing the behavior of language models, researchers often resort to probing: searching for connections between the behavior of neural network elements and properties of processed data. The aim of such studies is to establish the level at which the model internalizes specific levels of language (e.g., morphology, syntax, or discourse in a hierarchical model of language levels [27]). One possible direction of such research is to analyze how linguistic structures represent the syntagmatic and paradigmatic mechanisms of language in the models' knowledge. Sometimes [28], researchers presume that the ability to identify such structures indicates the models' approximation to human-level expertise in practical tasks.

Putting in order various studies on probing structural interpretation of language models, [29] distinguishes three stages (most works correspond to the first two stages) in the development of probing research: "behavioral," "correlational," and "invasive" probing. These paradigms naturally progress in stages.

In the behavioral probing stage, researchers establish the fact of a language model's awareness of a certain property. During the correlational probing stage, specific regions — groups of weights (often entire layers of weights) — whose behavior notably correlates with the investigated property are identified. For instance, in the

work by [30], authors identified neurons in a computer vision model that exhibited high activation specifically when processing images of heads.

During the invasive probing stage, the identified regions undergo more in-depth investigation: the weights identified by correlational probing are modified, and the behavior of the resulting model is compared to the behavior of the original model. This approach provides a visual representation of the relationship between individual elements of the language model and the properties of the processed data but is largely complementary to correlational research.

Probing tasks enable the measurement of linguistic awareness in model regions such as layers [31] or groups of neurons [32]. Measurement occurs as follows: an external model attempts to establish a connection between a pre-annotated dataset and records of the behavior of a certain part of the neural network model on it. The easier it is to establish such a connection, the more the behavior of the investigated part of the neural network correlates with the annotated property in the data. Usually, the external model acts as a classification model, reconstructing annotations in the text based on the behavior of the neural network model. In this case, the researcher's attention is directed at the performance quality of this external model. The external model is called the probing model, and the process of conducting such experiments is termed probing. Datasets [33], accompanied by linguistic annotations for probing language data, are termed probing datasets.

Existing probing datasets cover a wide range of linguistic features in texts, such as token parts-of-speech and syntactic parse trees. Thus, probing experiments suggest that the sought-after structure or linguistic property is indeed encapsulated within specific regions of the model. This allows further characterization of language models' ability to solve tasks related to language and linguistics specifically [34—37]. Studies relying on probing methods that indicate linguistic specializations in neural network segments include works like [31; 33; 38—44]. In these studies, researchers aimed to determine how semantic and grammatical knowledge is encoded in language models.

Syntactic and morphological probing involve tasks related to establishing the grammatical structure of texts based on vector representations generated by the model. Semantic probing tasks involve reconstructing word meanings, semantic relationships, or facts from a knowledge base using vector representations.

Early methods in this field, relying on part-of-speech and morphologically annotated probing datasets [38], enabled the establishment of what syntactic information can be encoded in layers of language-processing neural network models. Later methods allow for a more explicit determination of whether it's possible to reconstruct entire syntactic trees of texts based on the internal representations of pre-trained models.

For instance, in [43], a structural probing method was introduced to establish hierarchical tree-like structures from language vector representations. The results of applying this method to various large transformer models suggest that certain intermediate computations in such models are associated with the extraction of hierarchical structural properties in texts. (This general formulation is used due to the lack of explicit correlation with syntax theory.)

A separate line of research focuses on probing experiments with models in different languages. In [45], syntactic hierarchical generalizations made by models in various languages were identified using the structural probing method. In another work, [46], a methodological framework for multilingual morphosyntactic probing comprising 15 distinct tasks aimed at different languages was proposed. Studies based on this framework revealed that cross-linguistic typological regularities can be identified through probing experiments.

The probing methodology has also been applied to determine semantic information in vector representations produced by individual model components. Methods based on reconstructing semantic annotations from context illustrate the ability of contextual representations to disambiguate multiple meanings of a word represented by a single semantic vector [47]. Using a similar method, [39] conducted a layer-wise experiment on semantic class annotation recovery, revealing that higher layers of models outperform others in solving semantic tasks. However, probing relationships proposed in [31] indicated that semantic tasks yield lower probing task results compared to grammatical ones. The attention mechanism in today's popular transformer neural network architecture weighs all possible word-to-word connections in a sentence, computes weights, and uses them for text vectorization and subsequent language modeling task resolution. The differentiation of semantic and syntactic influences on computing these connection weights is challenging due to the scarcity of works on semantic probing tasks [21; 48]. Hereafter, we describe an approach to semantic probing designed to address such imbalance.

1.2.2 Interpreting Knowledge Representations in Language Models

In this work, we employ semantic probing based on the annotation of semantic relations. Semantic relations and the associated terms describe extralinguistic reality. Such encyclopedic or ontological descriptions form the knowledge investigated in our work using probing methods. Semantic probing tasks formulated based on knowledge graphs (tasks related to semantic relations and their participants) bear similarity to syntactic probing tasks (tasks related to predicates and their dependents).

Before conducting semantic probing, experiments [18] were conducted to justify the appropriateness of a layered language model in interpreting models. Subsequent research aligns with the behavioral and correlational stages of the three-stage probing investigation. In [19], behavioral probing aimed to establish the models' ability to correctly process relations presented in texts. However, the dataset formed at this stage proved excessively granular for conducting correlational probing. In [20], a more extensive dataset based on a knowledge graph is formulated, representing simpler semantic regularities.

The tools developed for experiments in [20], [19], and [18] formed the basis of a framework allowing researchers to conduct diverse probing studies involving relations in neural networks.

Justification of the Appropriateness of the Layered Language Model in Probing Research

In the study by [21], probing was conducted within the context of a layered language model. Correlational probing tasks representing morphological, syntactic, and discourse levels of language were posed to models at different stages of their training. This chronological analysis involved two models: BERT and T5, corresponding to encoder and encoder-decoder architectures, respectively. Twelve probing tasks were adopted from existing works—SentEval [33], Morph Call [18], DisSent [49], DiscoEval [50], and BLiMP [51]—to ensure result comparability:

- Subject number (SentEval): Grammatical probing task for binary classification on the subject's number.
- Person (Morph Call): Grammatical probing task for binary classification on the presence of person markers.
- Tree depth (SentEval): Grammatical probing task on the depth of parsing trees formulated in terms of classification.
- Top constituents (SentEval): Grammatical probing task to determine the main constituent in a sentence.
- Connectors (DisSent): Discourse probing task on restoring missing prepositions between pairs of sentences.
- Sentence position (DiscoEval): Discourse probing task on correctly ordering a sentence among four other sentences.
- Penn Discourse Treebank (DiscoEval): Discourse probing task on reconstructing discourse relations between Penn Treebank entities.
- Discourse coherence (DiscoEval): Discourse probing task to determine paragraphs with mixed sentences.

Thus, analyzing the quality of solving probing tasks allows drawing conclusions in favor of the appropriateness of a hierarchical language model in probing the language models. An explicit separation between morphology and syntax cannot be achieved, but grammatical and discourse knowledge of the models are consistently differentiated. The models show similar results in solving tasks despite differences in architectures and achieve final quality in grammatical probing tasks at the very beginning of training: over the first 100,000 iterations. Subsequently, the quality of solving these tasks remains at the achieved level. Discourse tasks prove to be much more challenging for the models: their quality steadily increases throughout the observed training period. The Adjunct Island and Top constituents tasks are poorly designed, and the models' unstable behavior on them might indicate an inappropriate choice or quantity of data.

Grammatical levels of the language (syntax and morphology) turned out to be difficult to distinguish through probing language models, while the separation between grammatical and ungrammatical levels is evident. Further, we explore both the interaction between semantics and grammar and consider semantics in an extralinguistic context.

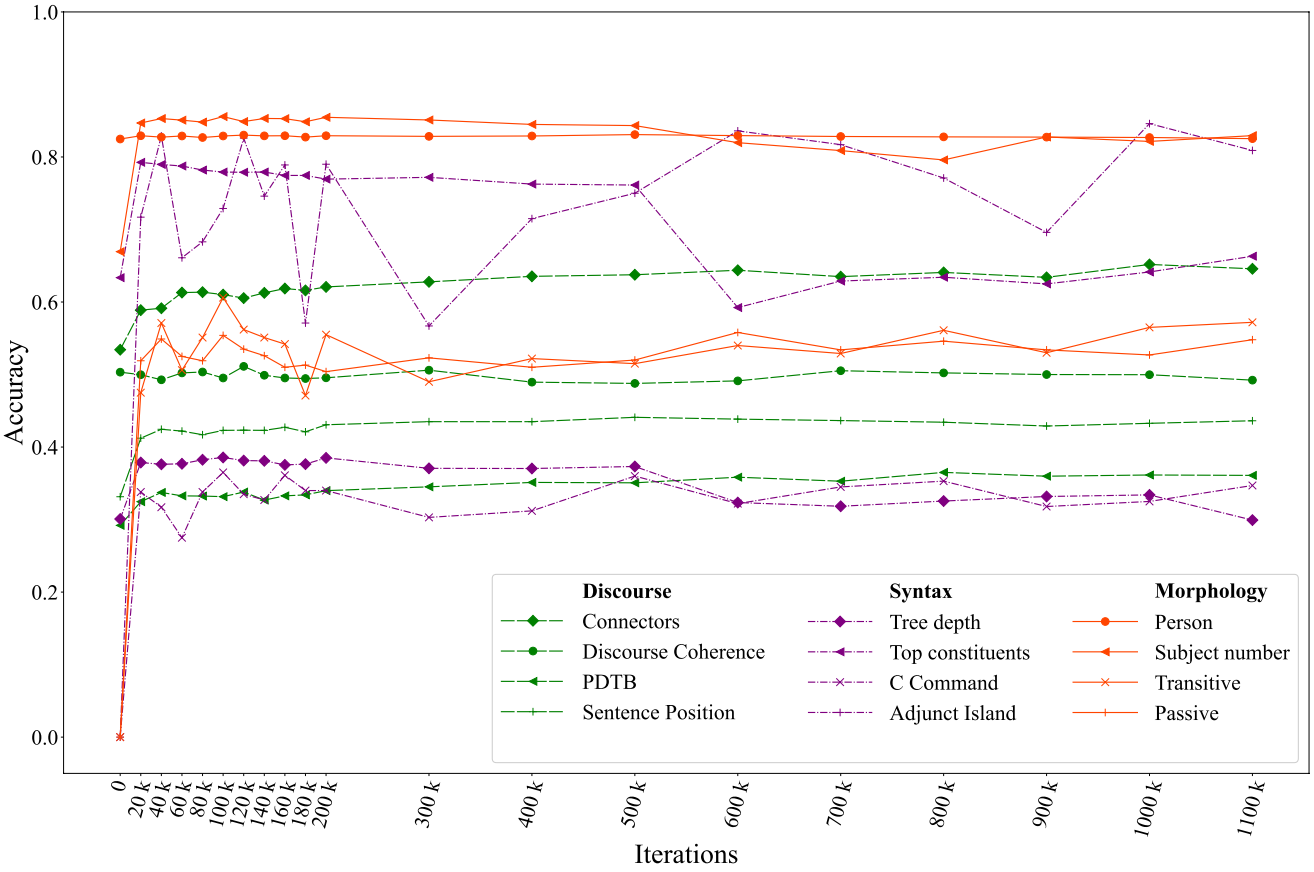

Рис. 1.1 — Results of experiments on BERT's acquisition of language levels. The figure displays the quality of solving probing tasks by models at different stages of training these models. Grammatical tasks are solved well starting from approximately 100,000 iterations, while the quality of solving discourse tasks gradually improves throughout the entire training of the models.

Within the scope of the work [19], a competition was conducted where participants were tasked with automatically assigning sentences to semantic sketches using language models. Semantic sketches (hereafter "sketches") are generalized lexicographic portraits where dependent words are classified based on their grammatical relationships: for dependent children, it is indicated whether they serve as a specifier, subject, object, or correspond to another syntactic role. Representing such information in a table format with popular placeholders for syntactic roles allows the construction of a semantic portrait of the corresponding predicate.

Sketches allow for a more precise investigation of the contribution of non-core arguments to the predicate. For instance, the Locative, a typical circumstantial adjunct, is an obligatory participant for verbs with spatial meanings, such as 'to be located'. It's the locative that enables the differentiation of various meanings of the verb 'to be', while the subject proves to be much less informative for conducting such differentiation. Semantic sketches are illustrated in Figure 1.2.

Conducting research to establish the models' ability to relate sentences to sketches corresponding to their head predicates allows us to assess whether the models possess the capability to consider semantics in the context of word relationships. The baseline solution provided to participants matched sentences and sketches by searching for words dependent on the sentence predicate in the sketch tables.

This baseline solution, along with subsequent solutions proposed by participants, was based on the RuBERT language model [52].

Baseline Solution Algorithm For a given context *ctx*, the sketch *sketch* was selected based on ranking using the following masking mechanism:

1. Initially, syntactic analysis was performed using the UDPipe solution ([53]) to find elements directly dependent on the main predicate in the sentence.
2. For each dependent element, the top N candidates for replacement Rep^N_{dep} were remembered.
3. Candidates selected in the previous step were intersected by lemmas $MLM^N_{ctx} = \bigcap_{dep \in ctx} Rep^N_{dep}$.

State	Relation Correlative	Relation Relative	Object	Object Situation	Time
хороший было хорошо	человек был человек	это был это	человек был он	это было это	тогда было тогда
так было так	поэт быть поэтом	то есть то	лицо было лицо	оно был он	время был в то время
нужный было нужно	друг были друзьями	отец был он	отец был отец	дело было дело	год был в те годы
готовый был готов	то был то	человек был он	время было время	жить была жизнь	раньше было раньше
трудный было трудно	ребенок был ребенком	Но есть но	глаз были глаза	голос был голос	теперь был теперь
невозможность было невозможно	женщина быть женщиной	жить есть жизнь	женщина была она	разговаривать был разговор	жить был всю жизнь

Object Situation	Time	Object Situation Like	Locative	Modality	Time AlreadyStill
это было это	потом было потом	день был день	часы было на часах	на самом деле было на самом деле	уже было уже
что было что	вчера было вчера	вечер был вечер	исход были на исходе	нормальный будет нормально	еще было еще
дело было дело	давний было давно	ночь была ночь	двор был на дворе	наяву было наяву	все было все
оно был он	завтра будет завтра	утро было утро	тут было не тут-то	возможный было возможно	по прежнему было по-прежнему
то было то	год было в проплом году	час было два часа ночи	гора была не за горами	невозможность было невозможно	все еще было все еще
случай был случай	поздний было позднее	весна была весна	СССР было в ссср	наверно было наверное	

Рис. 1.2 — Semantic sketches corresponding to non-locative and locative meanings of the verb ’to be’

4. The score *Score* of the sketch was calculated as the number of tokens that appeared in the intersection of the sketch representation with the remembered replacements:

$$Score\,(sketch, ctx) = \left| MLM_{ctx}^{1000} \cap Tokens_{sketch} \right|$$

Participant Solutions Analysis The systems participating in the competition were based on approaches that required retraining language models to solve tasks, as well as, particularly relevant from the perspective of interpreting language models, systems based on using systems without any training.

One of the participant systems disregarded the structure of sketches, naively converting the sketch table into a sequence, vectorizing such text to obtain a sketch vector. The sketch vectors were compared to sentence vectors based on cosine similarity between them.

Another system, surpassing the former, proposed an approach based on generating predicates corresponding to their sketches. To generate such hypotheses, templates motivated by the structure of semantic roles in predicates were used.

For the given semantic sketches, the concealed predicate was initially reconstructed. For this purpose, a technique borrowed from classical methods of resolving homonymy was used — template-based generation. An example of such a template might be "[MASK] to school,"suggesting the verb "go"as a probable predicate in place of the mask. By matching the obtained predicates with sentences, the relationship between sentences and sketches itself was reconstructed.

Thus, it appears that models do indeed pay attention to semantic relationships. However, the relatively modest improvement in quality suggests that semantic connections are highlighted by models in ways significantly different from how they are represented in annotated data sources.

Probing Models for Ontologies

The article [20] presents experiments and analytical insights from a correlational probing study of language models using BERT as an example. The work

draws inspiration from the findings of [54], an approach to automatically assigning words to ontologies. The authors of [54] demonstrate the ability of large language models to solve the problem of homonym disambiguation by referring to data sourced from wikidata [55].

To establish how relational semantic and factual knowledge is embedded within a language model, an approach comprising multiple stages was proposed. In the first stage, an algorithm was developed to perform relation extraction in texts. The algorithm intentionally relies on straightforward approaches and involves sequentially solving two classification tasks: initially filtering from all possible triples in the text those that potentially represent a semantically correct triplet <subject-predicate-object>, and then attributing the selected triples to specific types of semantic relations. Both layer activations of the model and activations of the attention heads forming these layers were experimented with for vectorizing token triples.

The analysis of the components of the described system allowed determining the degree of specialization of each attention head on each type of relationship: earlier layers of the model were found to be more informed about semantic relationships. By observing the behavior of the inter-token attention mechanism, it might be possible to ascertain whether the tokens under consideration are involved in any semantic relationship and even identify the nature of that relationship. Attention heads do not exhibit specialization for specific types of relationships, providing varying degrees of informativeness in establishing each relationship type. However, it is precisely this difference in the degrees of informativeness of attention heads that allows grouping relationship types. Representing such grouping as a tree via agglomerative clustering (Figure 1.2) indicates a trend towards merging similar relationship types. An interesting observation is that the hierarchy of relationships presented in the Wikidata resource differs from the one computed using the deterministic approach described. Nonetheless, the use of information from the computed hierarchical connections for the task of identifying semantic relationships in text proves to be more useful than providing information about relationships based on Wikidata.

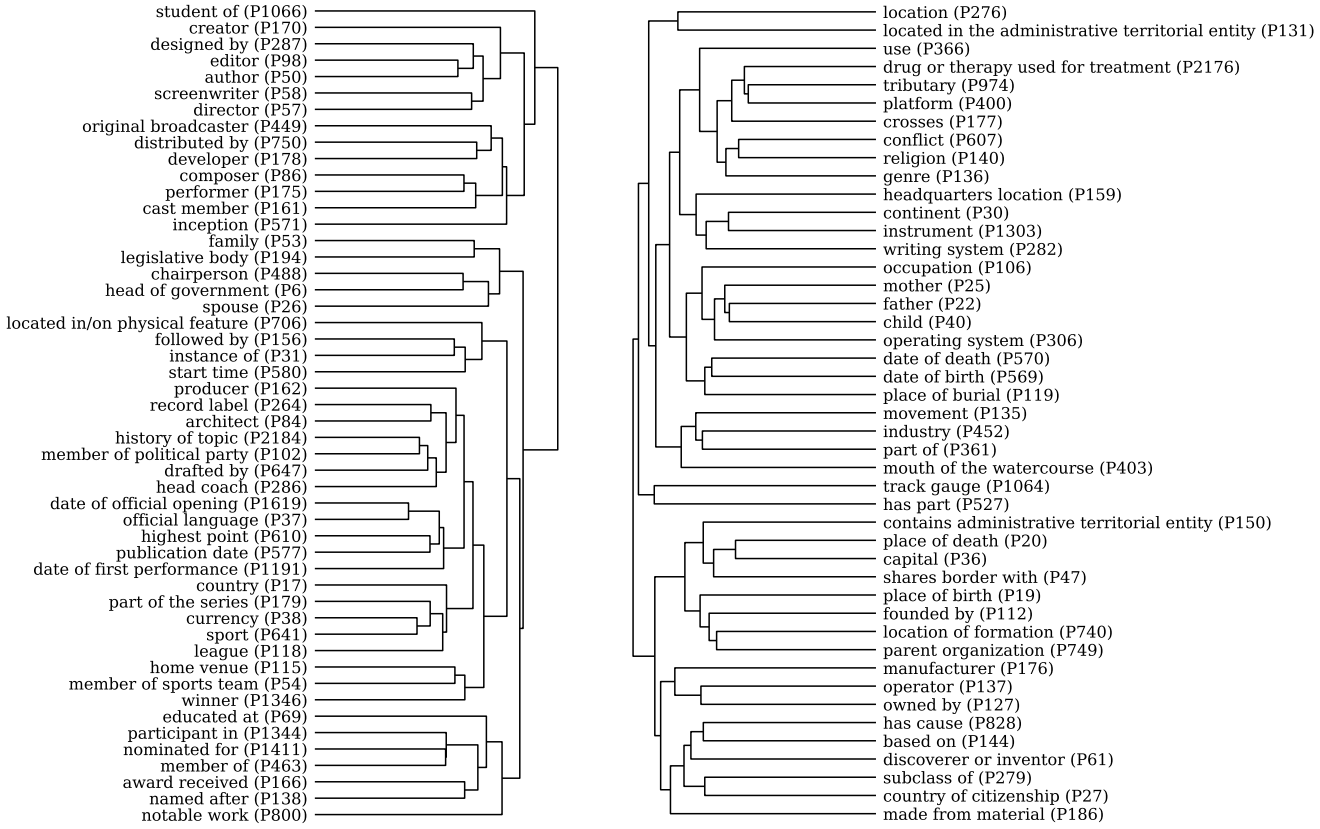

Рис. 1.3 — Agglomerative clustering of relationships. Through probing the BERT model to determine semantic relationships, areas of the model especially crucial for defining each relationship were identified. Comparative analysis of these crucial areas for each relationship allowed grouping relationships based on the similarity of their critical areas. The resulting grouping is represented in the figure as an agglomerative clustering of relationships. Semantically similar relationships are determined to fall into the same or closely related clusters (e.g., relationships like "founded by,location of formation,place of birth").

1.3 Conclusion

This dissertation research presents, applies, and makes accessible the methodology of probing knowledge in neural network language models. One of the side effects of developing this methodology was the emergence and widespread availability of tools that enable the probing of language models.

The probing study of knowledge (semantic structures and relationships) was conducted in two stages. In the behavioral probing stage, the fundamental importance of various semantic roles for language models was identified, corresponding to these roles as the models are contextual. This investigation was carried out using a set of semantic sketches accompanied by lists of typical fillers for semantic roles. This allowed us to confirm that if the properties of texts are described in terms of relationships and their participants, a connection between these text properties and model behavior can be observed.

During the correlational probing stage, direct interpretation of knowledge representation in models was performed. Instead of semantic sketches, named relations from the ontological model WikiData were used. Language models were examined for the presence of structures responsible for facts outlined in the framework of two-place semantic predicates. It was shown that segments with selective specialization are absent at both the level of model layers and at a more detailed level of attention heads constituting these layers. Analyzing the attention head specialization of language models on types of WikiData predicates allows for organizing a hierarchy of these predicates systematically and points out the inconsistency in the existing subjective hierarchy.

The toolkit underlying the methodology of the conducted research has proven convenient for conducting other studies on interpreting the behavior of language models. It has been presented as a system for interpreting language models, detailed in our published work [22]. This system enables researchers to conduct a large number of experiments (hundreds and thousands) on interpreting language models and analyzing their results, all presented in a unified format and interface. The system was applied in interpreting a new language model called BLOOM (the experiments are detailed in our published work [14]), helping establish the limits of generalizing grammatical abilities of a language model to unfamiliar languages.

Existing large language models serve as the foundation for an increasing number of practical systems and research endeavors. However, the practical application and research of these models are hindered by the inability to explain the exact reasons behind the results generated by such models. One technique to organize such a degree of uncertainty is through probing interpretation in the context of known structural descriptions of the models' application domain. However, structural descriptions of knowledge have not previously been used in the context of model interpretation. The proposed methodology of probing studies, as developed and tested in this work, provides a theoretical foundation and toolkit for future research.

References

1. RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora / D. Dementieva [и др.]. —.
2. RuArg-2022: Argument Mining Evaluation / E. Kotelnikov [и др.] // arXiv preprint arXiv:2206.09249. — 2022.
3. Superglue: Learning feature matching with graph neural networks / P.-E. Sarlin [и др.] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — с. 4938—4947.
4. RussianSuperGLUE: A Russian language understanding evaluation benchmark / T. Shavrina [и др.] // arXiv preprint arXiv:2010.15925. — 2020.
5. *Блог Яндекса*. Алгоритм «Палех»: как нейронные сети помогают поиску Яндекса. — доступ 3 ноября 2023. — https://yandex.ru/blog/company/algoritm-palekh-kak-neyronnye-seti-pomogayut-poisku-yandeksa.
6. Conversational ai: The science behind the alexa prize / A. Ram [и др.] // arXiv preprint arXiv:1801.03604. — 2018.
7. Language models are unsupervised multitask learners / A. Radford [и др.] // OpenAI blog. — 2019. — т. 1, № 8. — с. 9.
8. Language Models are Few-Shot Learners / T. Brown [и др.] // Advances in Neural Information Processing Systems. т. 33 / под ред. H. Larochelle [и др.]. — Curran Associates, Inc., 2020. — с. 1877—1901. — URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
9. Training language models to follow instructions with human feedback / L. Ouyang [и др.] // arXiv preprint arXiv:2203.02155. — 2022.
10. *OpenAI*. GPT-4 Technical Report. — 2023. — arXiv: 2303.08774 [cs.CL].
11. *SberDevices*. Русскоязычная нейросеть от Сбера. — доступ 3 ноября 2023. — https://developers.sber.ru/portal/products/gigachat.
12. Language is not all you need: Aligning perception with language models / S. Huang [и др.] // arXiv preprint arXiv:2302.14045. — 2023.

13. *AIRI.* RUDOLPH: One Hyper-Tasking Transformer can be creative as DALL-E and GPT-3 and smart as CLIP. — 2022. — https://github.com/ai-forever/rudolph.

14. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model / B. Workshop [и др.]. — 2023. — arXiv: 2211.05100 [cs.CL].

15. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing / под ред. Y. Goldberg, Z. Kozareva, Y. Zhang. — Abu Dhabi, United Arab Emirates : Association for Computational Linguistics, 12.2022. — с. 11689—11698. — URL: https://aclanthology.org/2022.emnlp-main.803.

16. Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP / под ред. J. Bastings [и др.]. — Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics, 12.2022. — URL: https://aclanthology.org/2022.blackboxnlp-1.0.

17. Emergent structures and training dynamics in large language models / R. Teehan [и др.] // Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models. — 2022. — с. 146—159.

18. *Mikhailov V.*, *Serikov O.*, *Artemova E.* Morph Call: Probing Morphosyntactic Content of Multilingual Transformers // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — Online : Association for Computational Linguistics, 06.2021. — с. 97—121. — DOI: 10.18653/v1/2021.sigtyp-1.10. — URL: https://aclanthology.org/2021.sigtyp-1.10.

19. SemSketches2021: Experimenting with the machine processing of the pilot semantic sketches corpus | SemSketches2021: опыт автоматической обработки пилотного корпуса семантических скетчей //. т. 20. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2021. — с. 560—570.

20. Attention Understands Semantic Relations / A. Chizhikova [и др.] // Proceedings of the Thirteenth Language Resources and Evaluation Conference. — Marseille, France : European Language Resources Association, 06.2022. — с. 4040—4050. — URL: https://aclanthology.org/2022.lrec-1.430.

21. Is neural language acquisition similar to natural? A chronological probing study | Усвоение языка у языковых моделей и человека: хронологическое пробинг-исследование //. т. 21. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2022. — с. 550—563.

22. Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation / O. Serikov [и др.] // Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. — Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics, 12.2022. — с. 441—456. — URL: https://aclanthology.org/2022.blackboxnlp-1.37.

23. On the Opportunities and Risks of Foundation Models / R. Bommasani [и др.] // CoRR. — 2021. — т. abs/2108.07258. — arXiv: 2108.07258. — URL: https://arxiv.org/abs/2108.07258.

24. Self-Organization and Artificial Life / C. Gershenson [и др.] // Artificial Life. — 2020. — сент. — т. 26, № 3. — с. 391—408. — DOI: 10.1162/artl_a_00324. — eprint: https://direct.mit.edu/artl/article-pdf/26/3/391/1896088/artl_a_00324.pdf. — URL: https://doi.org/10.1162/artl%5C_a%5C_00324.

25. *Dresp-Langley B.* Seven Properties of Self-Organization in the Human Brain // Big Data and Cognitive Computing. — 2020. — т. 4, № 2. — DOI: 10.3390/bdcc4020010. — URL: https://www.mdpi.com/2504-2289/4/2/10.

26. *Gordon D. M.* The ecology of collective behavior in ants // Annual review of entomology. — 2019. — т. 64. — с. 35—50.

27. *Dalrymple M.* Lexical functional grammar. — Brill, 2001.

28. *McCoy R. T.*, *Min J.*, *Linzen T.* BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. — Online : Association for Computational Linguistics, 11.2020. — с. 217—227. — DOI: 10.18653/v1/2020.blackboxnlp-1.21. — URL: https://aclanthology.org/2020.blackboxnlp-1.21.

29. Probing for the Usage of Grammatical Number / K. Lasri [и др.] // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Dublin, Ireland : Association for Computational Linguistics, 05.2022. — с. 8818—8831. — DOI: 10.18653/v1/2022.acl-long.603. — URL: https://aclanthology.org/2022.acl-long.603.

30. An overview of early vision in inceptionv1 / C. Olah [и др.] // Distill. — 2020. — т. 5, № 4. — e00024—002.

31. What do you learn from context? Probing for sentence structure in contextualized word representations / I. Tenney [и др.] // arXiv e-prints. — 2019. — май. — arXiv:1905.06316. — arXiv: 1905.06316 [cs.CL].

32. Analyzing individual neurons in pre-trained language models / N. Durrani [и др.] // arXiv preprint arXiv:2010.02695. — 2020.

33. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties / A. Conneau [и др.] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 07.2018. — с. 2126—2136. — DOI: 10.18653/v1/P18-1198. — URL: https://www.aclweb.org/anthology/P18-1198.

34. *Kitaev N.*, *Klein D.* Constituency Parsing with a Self-Attentive Encoder // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 07.2018. — с. 2676—2686. — DOI: 10.18653/v1/P18-1249. — URL: https://aclanthology.org/P18-1249.

35. Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling / L. He [и др.] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Melbourne, Australia : Association for Computational Linguistics, 07.2018. — с. 364—369. — DOI: 10.18653/v1/P18-2058. — URL: https://aclanthology.org/P18-2058.

36. Linguistically-Informed Self-Attention for Semantic Role Labeling / E. Strubell [и др.] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — Brussels, Belgium : Association for Computational

Linguistics, 11.2018. — c. 5027—5038. — DOI: 10.18653/v1/D18-1548. — URL: https://aclanthology.org/D18-1548.

37. *Lee K.*, *He L.*, *Zettlemoyer L.* Higher-Order Coreference Resolution with Coarse-to-Fine Inference // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). — New Orleans, Louisiana : Association for Computational Linguistics, 06.2018. — c. 687—692. — DOI: 10.18653/v1/N18-2108. — URL: https://aclanthology.org/N18-2108.

38. What do Neural Machine Translation Models Learn about Morphology? / Y. Belinkov [и др.] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Vancouver, Canada : Association for Computational Linguistics, 07.2017. — c. 861—872. — DOI: 10.18653/v1/P17-1080. — URL: https://aclanthology.org/P17-1080.

39. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks / Y. Belinkov [и др.] // Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — Taipei, Taiwan : Asian Federation of Natural Language Processing, 11.2017. — c. 1—10. — URL: https://aclanthology.org/I17-1001.

40. Dissecting Contextual Word Embeddings: Architecture and Representation / M. E. Peters [и др.] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — Brussels, Belgium : Association for Computational Linguistics, 10.2018. — c. 1499—1509. — DOI: 10.18653/v1/D18-1179. — URL: https://aclanthology.org/D18-1179.

41. *Zhang K.*, *Bowman S.* Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. — Brussels, Belgium : Association for Computational Linguistics, 11.2018. — c. 359—361. — DOI: 10.18653/v1/W18-5448. — URL: https://aclanthology.org/W18-5448.

42. *Alain G.*, *Bengio Y.* Understanding intermediate layers using linear classifier probes // 5th International Conference on Learning Representations, ICLR

2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. — OpenReview.net, 2017. — URL: https://openreview.net/forum?id=HJ4-rAVtl.

43. *Hewitt J.*, *Manning C. D.* A Structural Probe for Finding Syntax in Word Representations // 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). — Minneapolis, Minnesota : Association for Computational Linguistics, 06.2019. — с. 4129—4138. — DOI: 10.18653/v1/N19-1419. — URL: https://aclanthology.org/N19-1419.

44. *Hewitt J.*, *Liang P.* Designing and Interpreting Probes with Control Tasks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China : Association for Computational Linguistics, 11.2019. — с. 2733—2743. — DOI: 10.18653/v1/D19-1275. — URL: https://aclanthology.org/D19-1275.

45. *Chi E. A.*, *Hewitt J.*, *Manning C. D.* Finding Universal Grammatical Relations in Multilingual BERT // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 07.2020. — с. 5564—5577. — DOI: 10.18653/v1/2020.acl-main.493. — URL: https://aclanthology.org/2020.acl-main.493.

46. Linspector: Multilingual probing tasks for word representations / G. G. Şahin [и др.] // Computational Linguistics. — 2020. — т. 46, № 2. — с. 335—385.

47. Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings / Y. Yaghoobzadeh [и др.] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 07.2019. — с. 5740—5753. — DOI: 10.18653/v1/P19-1574. — URL: https://aclanthology.org/P19-1574.

48. *Rogers A.*, *Kovaleva O.*, *Rumshisky A.* A primer in bertology: What we know about how bert works // Transactions of the Association for Computational Linguistics. — 2020. — т. 8. — с. 842—866.

49. *Nie A.*, *Bennett E.*, *Goodman N.* DisSent: Learning sentence representations from explicit discourse relations // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — с. 4497—4510.

50. *Chen M.*, *Chu Z.*, *Gimpel K.* Evaluation benchmarks and learning criteria for discourse-aware sentence representations // arXiv preprint arXiv:1909.00142. — 2019.

51. BLiMP: The benchmark of linguistic minimal pairs for English / A. Warstadt [и др.] // Transactions of the Association for Computational Linguistics. — 2020. — т. 8. — с. 377—392.

52. *Kuratov Y.*, *Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. — 2019. — arXiv: 1905.07213 [cs.CL].

53. *Straka M.*, *Straková J.* Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe // Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. — 2017. — с. 88—99.

54. *Loureiro D.*, *Jorge A.* Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 07.2019. — с. 5682—5691. — DOI: 10.18653/v1/P19-1569. — URL: https://aclanthology.org/P19-1569.

55. *Vrandečić D.*, *Krötzsch M.* Wikidata: a free collaborative knowledgebase // Communications of the ACM. — 2014. — т. 57, № 10. — с. 78—85.