National Research University Higher School of Economics

as a manuscript

Tatiana Shavrina

LINGUISTIC INTERPRETATION AND EVALUATION OF THE WORD VECTOR MODELS FOR RUSSIAN

Dissertation Summary
for the purpose of obtaining
academic degree Doctor of Philosophy in Philology and Linguistics

Academic Supervisor:
Olga Lyashevskaya, PhD

Moscow 2022

The dissertation was prepared at the National Research University "Higher School of Economics."

Publications

The five articles listed below are submitted for defence, in two of which the applicant is the sole author, in two — the first author, and in one — the executor in the project.

1. **Shavrina T. O**. Methods of computational linguistics in the evaluation of artificial intelligence systems. Voprosy Jazykoznaniya, 2021. № 6. P.117-138.
   Citation database: Q2 Scopus

2. **Shavrina T.** Word vector models as an object of linguistic research. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". 2019. P. 576-588. Citation database: Scopus

3. Alena Fenogenova, **Tatiana Shavrina**, Alexandr Kukushkin, Maria Tikhonova, Anton Emelyanov, Valentin Malykh, Vladislav Mikhailov, Denis Shevelev, Ekaterina Artemova. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021". 2021. P .235-246. Citation database: Scopus

4. **Shavrina T.**, Fenogenova A., Emelyanov A., Shevelev D., Artemova E., Malykh V., Mikhailov V., Tikhonova M., Chertok A., Evlampiev A. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2020. P. 4717-4726. Citation database: CORE A (Computer Science)

5. **Shavrina T.,** Emelyanov A., Fenogenova A., Fomin V., Mikhailov V., Evlampiev A., Malykh V., Larin V., Natekin A., Vatulin A., Romov P., Anastasiev D., Zinov N., Chertok A. Humans Keep It One Hundred: an Overview of AI Journey. Proceedings of The 12th Language Resources and Evaluation Conference Vol. 12. European Language Resources Association (ELRA). 2020. P. 2276-2284. Citation database: Scopus

Conference presentations and public demonstrations of the results

The main results and conclusions of the present study have been presented in 2019–2021 in oral presentations at nine international conferences:

1. DIALOGUE 2021 conference, 16 June 2021

Report: Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models link online, Moscow, Russia

2. DIALOGUE 2019 conference, 29 May 2019. Report: Word vector models as an object of linguistic research link , Moscow, Russia

3. Artificial General Intelligence conference, AGI-2020,
   13th International Conference, AGI 2020, St. Petersburg, Russia, September 16–19, 2020
   Report: Russian SuperGLUE Creating a Language Understanding Evaluation Benchmark link

4. The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020, 16 – 20 November 2020. Report: RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark link online, Dominican Republic

5. Symposium on Big Data Analysis to Identify Global Challenges and Trends in Human Development. Session 3 in the framework of the XXII April International Scientific Conference of the Higher School of Economics
   12 April 2021, Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics. Report: «All ways to measure an elephant: Russian Superglue & RuSentEval» link online, Moscow, Russia

6. XIII Shmelev Readings (EVERYDAY SPEECH AS AN OBJECT OF LEXICOGRAPHY), 23–25 February 2020, Moscow, Volkhonka, 18/2
   Report: "Universal models on the corpus of everyday speech as new lexicography tools". RAS Institute of the Russian language of V.V. Vinogradov link

7. 4th Kolmogorov Seminar on Computational Linguistics and Language Sciences
   Report: Struggling with word vector models interpretation: some experience gained from basic linguistic practice
   Staraya Basmannaya st., 21/4, School of Linguistics, National Research University Higher School of Economics. link Moscow, Russia

8. Kostomarov forum, discussion "Does the computer speak and understand?" Laboratory for Cognitive and Linguistic Research, Pushkin Russian Language Institute, 25 May 2021 link online, Moscow, Russia

9. Moscow HSE Pragmatics Workshop. Tatiana Shavrina: Russian Commitment Bank: machine learning lessons vs lessons of linguistics -- all not learnt? 30 September 2021 link online, Moscow, Russia

1. Topic, contents and structure of the paper

The proposed dissertation focuses on methods for linguistic interpretation and evaluation of word vector models for the Russian language.

Word vector models occupy an essential place in the field of natural language processing (NLP) and are an integral basis for solving a wide range of problems, such as text classification (e.g. determining the topic of a text, analyzing the emotions of a text, classifying offensive messages), information extraction (e.g. recognition of named entities, extraction of facts), machine translation, text summarization, as well as text generation (automatic creation of unique texts of a given genre on a given topic). Various methods of statistics and machine learning, including neural networks, now result in vector models of words and texts.

Vector models work with words and texts in a vector space of features, associating a text or a word with a numerical vector of a fixed length. As stated by (Conneau A. et al. 2018), one of the main challenges of working with the models is the opacity of the "black box" of a vector model resulting from the way it is trained. The features presented in vectors of fixed length are poorly interpretable, which significantly complicates the comparison and selection of the best vector model from the available ones. This poses a great challenge for comparing models, their performance and interpretability, as formulated in the work (Rogers et al. 2020). The first steps in this direction are underway for the English language, primarily by developing a benchmark methodology (Wang A. et al. 2018, Wang A. et al. 2019). Due to the development of the engineering basis of language modelling at the moment there are hundreds of varieties of various vector models of words and texts, including most of them adapted for use in Russian-language material, for example, the output of RusVectores project (Kutuzov, Kuzmenko 2017) and DeepPavlov (Kuratov, Arkhipov, 2018).

Research Goals

The methodology design for the timely evaluation and interpretation of the vector model training allows one to simultaneously achieve two important goals that this study is devoted to following objectives.

- First, we aim to develop criteria for evaluating vector models — these criteria form the landscape of work that is carried out by the community in order to improve the current results of language modelling for years to come. The formulation of the modelling criteria that are more substantiated from the point of view of theoretical ideas about the language provides the basis for the development of both language engineering prototypes and tools for evaluating these prototypes.
- Secondly, our goal is to make the learning results of neural networks more understandable for humans and identify additional factors influencing the quality of language modelling. The vector model interpretation tools, in particular, allow shedding light on the “black box” of neural networks, and this makes the development of such tools for Russian-language models in demand.

To achieve the stated goals, we completed the following **tasks**:

- to overview and compare of the methods for evaluating vector models for various languages, identifying limitations in theoretical requirements for the results of language modelling and practical evaluation of the results;
- to overview the limitations of various vector model architectures for words and texts, including current models based on the transformer architecture and older models of distributional semantics (word2vec (Mikolov T. et al. 2013), GloVe (Pennington et al. 2013)), as well as basic vectorization models based on TF-IDF;
- to create a set of new text-based tests for evaluating the modelling of various linguistic intellectual abilities, including of tests for 1) conducting causal relationships between events in the texts, 2) natural language inference, 3) general and encyclopedic knowledge, commonsense, logic, as well as machine reading — the so-called benchmark for the Russian language, named Russian SuperGLUE;
- th create the "linguistic diagnostics": a set of diagnostic tests that determine the impact of various phenomena of morphology, syntax, lexical and formal semantics, world knowledge on the model training results ;
- to prepare a codebase that ensures the invariance of conducting tests with a model of any architecture (neural network, distributive-semantic, rule-based, etc.);
- to test the existing vector models of words and texts for the Russian language on the obtained evaluation and interpretation system, analyzing the results, measuring the average human level in solving the above problems.

The relevance of the study is determined by two main factors:

- the rapid development of neural language modelling provides new model artefacts without their profound evaluation or validation, that blocks the best solutions to be highlighted for further progress in natural language understanding;
- the lack of evaluation and interpretation systems for Russian makes it impossible to assess vector models of words and texts for the Russian language.

The author's contribution is determined by the following provisions: in (Shavrina T. 2019), the author single-handedly proposed a methodology for interpreting and comparing static vector models, developing a codebase and methods for testing the generalizing ability of models in the field of lexical semantics. In the work (Shavrina T. et al. 2020b), the author led the development of experimental software based on vector models, which completely solves the variants of the Unified State Exam in the Russian language, including tests, tasks with an open answer and an essay. The author's work included motivation and formulation of the problem, developing a methodology for the unified solution and also developing solutions for 6 types of questions. In the works

(Shavrina T. et al. 2020, Fenogenova A. et al. 2021), the author was responsible for the development of a methodology for evaluating and interpreting vector models, as well as collecting primary text data for subsequent filtering and editing in the subcorpora of the tasks. The work (T.O. Shavrina 2021) summarized the above experiments and combined them in an overview describing the methodological prerequisites, the motivation for the decisions made, and the current limitations of the proposed methodology.

Thus, within the framework of this study, the following provisions are submitted to the defence.

1) The gradual progress of vector models of words and texts is measured using a set of various intellectual tasks, providing objective fixed conditions that do not give advantages to any of the tested models.
2) The set of tasks for testing language modelling should include tasks that are complex enough for the current level of development of applied language technologies; such a difficult level is offered by the General Language Understanding Evaluation (GLUE, SuperGLUE) methodology.
3) Vector models of words and texts for the Russian language show the ability to identify correlations between various formulations of intellectual tasks and linguistic phenomena that are explicitly expressed lexically, for example, *to solve textual inference tasks better than a random choice, if the formulation contains negation, disjunction, conjunction or conditional construction.*
4) However, identifying these correlations is not enough to solve the tests without errors. None of the publicly presented vector models for the Russian language came close to the human level in solving the presented word problems. With the help of the benchmark presented in the work, significant errors and contradictions in language modelling, modelling of the vector space of words and texts for various models have been fixed.

The theoretical significance of the dissertation is determined by the general convergence of the achievements of linguistics and the theory of artificial intelligence, including the following factors:

- as the main tool for assessing the level of intelligence of systems, language tests were presented that assess the morphological, syntactic, semantic, pragmatic and discursive levels of the language.
- for the first time, a procedure for testing intelligent systems for the Russian language was compiled and described, including training, validation and testing procedures, as well as a detailed analysis of the results, diagnosis of errors and comparison with the human level.

The practical significance is determined by the introduction of a new toolkit, the Russian SuperGLUE rating, which consists of 9 new corpora of intellectual tests for the Russian language; each corpus of tests is divided into 3 fixed parts — a training sample, a sample for self-testing of the participants, and a test sample with closed golden answers. The toolkit is available online[1] under an open-source license. Since the public launch of online access to the rating (June 2020), 1530 different variations of vector models for the Russian language have been tested and interpreted; 22 of these models are publicly ranked[2] against the human level.

The impact of the provided research is presented as a set of theoretical and practical achievements in the framework of benchmarking and new text corpora.

2. General features of vector models of word and text

Vector models are capable of representing words and texts in the form of numerical features suitable for processing by various algorithms. The resulting feature vectors corresponding to a word or text can be used to determine words that are close in meaning, similar in subject matter, and can also be subjected to various mathematical operations (Turney, Pantel, 2010): for example, find a word A that is in the same relation to word B as a word C to D:

“Moscow”—”Russia”, “Seoul”—?
Answer:”South Korea”[3]

Vector models are conventionally divided into two categories:

- static, in which the vector of each word or text is strictly fixed and uniquely determined by the results of training the vector model on a certain corpus of texts; The disadvantages of such models include the coincidence of feature vectors for homonyms and polysemantic words, as well as random vectors for the most frequent words of the service parts of speech found in a wide variety of contexts;
- and dynamic, or contextual, in which the vector of features of a word or text depends and can vary significantly depending on the collocates on the left and right, being an indicator of the context value.

Models of the first type (static) include vector models such as

- simple collocation models, vector space models based on methods and corpus statistics. Models of this kind collect the frequencies of the co-occurrence of all unique words in the corpus: for example, the word "linguistics" appears in the same text with the word "computer" 200 times per 10 billion words, and "corpus" occurs in the same text with the word “linguistics” 300 times per 10 billion words. So, for each word, a vector of length with the size of the dictionary is collected, where each number corresponds to the frequency of occurrence of the word with each other. Such vectors, of course, contain many zero elements,

[1] https://russiansuperglue.com
[2] the publicity of the result in the rating is determined by the desire of the author of the system. The rating is presented at https://russiansuperglue.com/leaderboard/2
[3] Based on the word2vec vector model trained on the texts of the RNC and Wikipedia https://rusvectores.org/ru/calculator/#

and also have an extremely large dimension, since the number of unique entries in a large corpus dictionary can be equal to millions of words.

- neural models of distributive semantics: word2vec, fasttext, Glove and other models. Such models rely on simple collocation models, striving to efficiently compress vectors of large dimensions in various ways. In models of distributive semantics, primary frequencies of co-occurrence of words are often used not in the whole document, but in a small context, for example, at a distance of 5 words from each other. Efficient compression of large vectors occurs due to the neural network architectures Continuous bag of words (CBoW) or Skip-gram (Mikolov et al. 2013). CBoW is an architecture that learns to compress and decompress a word vector in a way that predicts a word based on its surrounding context. Skip-gram works the other way around: using the vector of the current word, the neural network learns to predict the surrounding words.

The second type of model, dynamic, is mainly formed by the so-called transformers: models based on the encoder-decoder architecture with the attention mechanism. The neural network encoder accepts text as input, and the attention mechanism weighs the importance of each word, setting the importance coefficients - based on them, the encoder forms the context vector, and the decoder solves the given problem - continues the text, or assigns some kind of classification label. Such architectures include, for example, the BERT (encoder only), GPT-3 (decoder only), T5 (encoder and decoder) models, and others.

3. Proposed methodology for evaluating and interpreting vector models

Static vector model evaluation

In (Shavrina, 2019) static vector models are considered as an independent object of linguistic research. Various static vector models of the Russian and English languages, their capabilities and disadvantages are considered in detail. It is concluded that with the help of statistical experiments on static vectors obtained in various corpora of the Russian language, stable vocabulary groups with the most homogeneous, stable contexts are distinguished, regardless of the genre and stylistic composition of the corpus. These vocabulary groups include adjectives denoting a person's personal qualities, nationality, profession, place names, adjectives of the time.

At the same time, proper names are the most unstable group, as they are the rarest and most context-sensitive. For the Russian language, an experiment was conducted to assess the residual amount of semantic and ontological relationships between known pairs of words, and the quality of the models was assessed based on this amount of relationships remaining in the model. It was found that words from the Swadesh list are more resistant to model change and retain their nearest vector neighbours much more often than words from the first thousand words of the frequency dictionary, and also more often than random words. These results are also reproduced for the English language.

At the same time, for quality analysis and interpretation of dynamic vector models, a different methodology is needed that is suitable for dynamic vectors - it is presented in the next section and is described in detail in (Shavrina T. et al. 2020a, Fenogenova A. et al. 2021).

Dynamic vector model evaluation

Since their appearance in 2016, dynamic vector models have been the technological basis for most applied solutions with state-of-the-art quality. With the help of dynamic vector models, for the first time on formal metrics, results were obtained above the average level of assessors: for example, in the problem of finding an answer to a question in Wikipedia (Stanford Question Answering Datasets SQuAD problem, for English (Li Yi, 2017)), on the news corpus the quality of human translation from Chinese into English is exceeded (Hassan H. et al., 2018)), and the level of human speech recognition (English) is also exceeded.

However, for this reason, standard applied tasks, such as finding answers to questions in a corpus, classifying texts by topics or sentiment, extracting named entities from text, and so on, are too simple for objective comparison of these models. Applied problems of processing the Russian language cannot provide a significant scatter of metrics between competing models and are often solved at a level equal to or higher than the level of an average human solution (95%+). In this case, the scatter of estimates between competing systems decreases, and their comparison becomes uninformative.

Since the advent of the Turing Test (Turing 1950), which provides an assessment of a machine's ability to imitate human intelligence in the messaging form between the AI and the judges, a wide range of related intelligence tests have emerged. These techniques are discussed in detail in (Shavrina, 2021). The practice of comparing the intellectual abilities of systems according to the results of one of these tests still dominates in the modern research community, however, to improve the reliability of the results, diversification of tests is required.

The approach that implements this strategy in evaluating intelligent systems is called benchmarking. It was first presented in (Fleming et al. 1986): comparison of computer systems in equal conditions requires accurate formulation of tasks and aggregation of results. The benchmark approach as applied to intelligent systems involves a combination of several principles:

1) Fixed data separation: a set of examples is collected for the formulated task, then is divided into three parts in a fixed way: a training sample, a sample for validation and a test sample for public comparison of systems (usually in a percentage ratio of 80-10-10% or 70-15 -15% of all examples).

2) Closedness of the test sample: “golden” answers to test tasks are inaccessible to participants and are not available for external search. The textual representation of intellectual problems allows the most diverse assessment of the abilities of the competing systems, including the tasks that require subject knowledge *(bees do not fly according to the same laws of physics as the plane does)*, commonsense knowledge about environmental objects and their interaction *(green fruits are not worth eating, yellow and red ones are already ripe)*, logic, the ability to establish causal relationships between the described events.

Russian General Language Understanding Evaluation

Both static and dynamic vector models demonstrate the ability to contribute to the solution of relatively simple problems with certain boundaries. Thus, in the work (Shavrina T. et al. 2020b) it is shown that with the help of static and dynamic vector models it is possible to assemble software for the automatic solution of the Unified State Exam in the Russian language, combining direct text sources of knowledge (texts of textbooks), statistical models for ranking answers, several models for the placement of punctuation, a neural network spell checking system, a system of rules for solving text comprehension tasks, a neural network model for generating the text of an essay. When working with the system, it becomes clear that the solution is not fully intellectual, since it only uses a fixed set of rules and facts, although it can demonstrate certain, quite satisfactory, results within the framework of the task. Neither each of its components, nor their totality, have knowledge of the Russian language, but on the whole, it demonstrates a level sufficient to simulate the successful completion of exam tasks - on average 69 points out of 100, which corresponds to the level of four points out of five.

If in simpler tasks, the word vector models demonstrate their superiority, then with highly intelligent tasks the situation is quite different. And for more complex intellectual tasks, a well-developed methodology is required to determine the degree of current levels of the problem-solving quality.

The General Language Understanding Evaluation (GLUE) methodology, first proposed for the English language, considers the evaluation of vector models in a complex: the model must demonstrate its level of solving intellectual text problems, preferably rather complex ones, simulating various abilities of a person: world knowledge, logic, common sense, the ability to conduct causal relationships, demonstrate the understanding of the text. This technique assesses the suitability of a model to solve a multitude of problems at once, and these problems themselves inherit the Turing test methodology: they include various textual formulations of questions, usually with multiple answers, and the model needs to “pretend to be human” - to choose the most correct answer.

For the Russian language, this method of interpretive evaluation of language models is being created for the first time and forms the basis of the Russian SuperGLUE project.

The project contains an updated rating of vector models of the Russian language, their evaluation based on their answers to questions, as well as the interpretation of the results based on model errors, and the correlation of errors with linguistic information of various levels - morphology, syntax, semantics, pragmatics.

Within the framework of the project, we created the following corpora of interpretive intellectual tasks for assessment of the suitability of models for the following tasks in Russian:

1. *Linguistic Diagnostic for Russian (LiDiRus):* establishing causal relationships on a corpus of minimal pairs of sentences with artificially complicated formulations and fixed linguistic properties of various levels.
2. *Russian Commitment Bank (RCB):* conducting causal relationships between events in news and fiction texts;
3. *Choice of Plausible Alternatives for Russian language (PARus):* making decisions based on common sense;
4. *Russian Multi-Sentence Reading Comprehension (MuSeRC):* establishing causal relationships in the read text;
5. *Textual Entailment Recognition for Russian (TERRa):* establishing causal relationships in comparable pairs of texts;
6. *Russian Words in Context (RUSSE)*, resolving semantic ambiguity based on context and common sense.
7. *The Russian Winograd Schema Challenge (RWSD):* solving logical problems and goal-setting;
8. *Yes/no Question Answering Dataset for the Russian (DaNetQA):* answering questions on subject knowledge and reading comprehension.
9. *Russian Reading Comprehension with Commonsense Reasoning(RuCoS):* reading comprehension;

The LiDiRus corpus (Problem 1) was added to the list, as it has a special purpose besides the classification task: linguistic interpretation. The linguistic interpretation of dynamic vector models implies the study of all kinds of dependencies between the learned vector features of words and texts and the known linguistic parameters, properties of training corpora. For this purpose, LiDiRus provides the testbed for the correlation analysis of the errors and linguistic features. The task can be solved with or without training the model, and generates an analytical report on the probabilities of the model to make mistakes depending on the following properties:

- Lexical Semantics: lexical entailment, factivity, quantifiers, named entities, symmetry or collectivity, morphological negation, redundancy;
- Logic: negation and double negation, intervals or numbers, upward/downward/non-monotone, temporal, conjunction and disjunction, conditionals, universal and existential;
- Predicate-Argument Structure: core arguments, prepositional phrases, intersectivity, restrictivity, anaphora and coreference, coordination scope, active or passive voice, ellipsis or implicits, nominalization, relative clauses, datives, genitives and partitives;
- Knowledge: common sense, world knowledge.

Examples of all categories are detailed in Appendix 1.

Various results for Russian model evaluation and interpretation

The Russian SuperGLUE methodology is both suitable for static and dynamic word vector model research.

By now, 1530 models have been evaluated with the benchmark, having their private record about the performance on various intellectual tasks and model's exposure to errors influenced by various language features. In table 1 you can see the top results for the Russian language model performance compared to the average human performance (by September 2021).

Table 1. Human level and the top 3 vector rating models based on the average score for 9 intellectual tasks. The overall score is calculated by averaging the results of each task. The specific task results use the following metrics: LidiRus - Matthews Correlation, RCB - F1 / Accuracy, PARus - Accuracy, MuSeRC - F1 / EM, TERRa - Accuracy, RUSSE - Accuracy, RWSD - Accuracy, DaNetQA - Accuracy, RuCoS - F1 / EM.

Название	**Overall score**	**LiDiRus**	**RCB**	**PARus**	**MuSeRC**	**TERRa**	**RUSSE**	**RWSD**	**DaNetQA**	**RuCoS**
1. human level	**0.811**	0.626	0.68 / 0.702	0.982	0.806 / 0.42	0.92	0.805	0.84	0.915	0.93 / 0.89
2.ruRoberta-large finetune	**0.684**	0.343	0.357 / 0.518	0.722	0.861 / 0.63	0.801	0.748	0.669	0.82	0.87 / 0.867
3. Golden Transformer	**0.679**	0	0.406 / 0.546	0.908	0.941 / 0.819	0.871	0.587	0.545	0.917	0.92 / 0.924
4.ruT5-large-finetune	**0.634**	0.32	0.306 / 0.498	0.66	0.815 / 0.537	0.747	0.735	0.669	0.711	0.81 / 0.764

At the moment, none of the existing models that have passed the testing have come close to the human level of solving intellectual problems: the total score of 81% of correct answers is quite far from the best modelling result (68% of correct answers, ruRoBERTa large model).

The rating of systems includes 22 different architectures, including dynamic vector models ruBERT, ruGPT-3, RuRoBERTa, their various variations and combinations.
The bottom lines of the rating are also represented by basic static solutions:

- a collocation model obtained with TF-IDF from the Wikipedia corpus,
- a random answer model,
- a model that always gives the same answer.

Among others, the ranking presents a rule-based solution based on heuristics: it takes 17th place. In general, the rule systems are at the bottom of the ranking, including solutions that always choose the same answer, as well as solutions based on static vector models: for example, a solution based on the TF-IDF model according to Wikipedia is ranked 20th line and 43.4% correct answers out of 100% possible. At the same time, such dynamic vector models as ruT5, ruRoberta, ruBert occupy the first positions in the ranking after the human level (68.6%, 63.5%, 62% of correct answers, respectively).

4. Conclusion

This study proposes a new methodology for assessing and interpreting the abilities of word vector models for the Russian language.

Word vector models show their excellent suitability for the best solutions in applied areas of computational linguistics, such as text classification problems, information extraction, machine translation. When considered as an independent object of research, they require special effort and evaluation to ensure the correct understanding of sentences and texts.

The proposed methodology suggests that one must put the “natural intelligence” of the human assessors and vector models in equal testing conditions for the more reliable modelling of the Russian language; the methodology includes:

- a set of 9 new corpora with tasks for various intellectual abilities, including those that separately measure the quality of subject knowledge, logic, cause-and-effect relationships, and comprehension of the text;
- a set of linguistic diagnostics, which checks the stability of the answers and their correctness, depending on the presence in the test of various phenomena of morphology, syntax, semantics;
- benchmarks of the human solution, solutions with popular word and text vector models, as well as heuristic models.

The gradual convergence of the methods of computational linguistics and general artificial intelligence is considered mutually beneficial: the theoretical understanding of the levels of the language allows you to create systems for testing and interpreting vector models, highlighting different levels of language acquisition by models: morphological, syntactic, level of lexical semantics, level of formal semantics, as well as levels of basic knowledge of language concepts.

Testing of intelligent systems based on textual tasks is a common method in the methodology for evaluating AI systems and has been developing since the 1960s, however, the property of linguistics began to be used in the formation of such tests relatively recently - with the advent of the general language understanding evaluation methodology. Within the framework of the proposed methodology, the quality of language modelling is assessed through a set of specialized skills expressed through language: possession of cause-and-effect relationships in the text, logical inference and disambiguation, decision-making within the described situations, operating with information about the basic properties and characteristics of objects of everyday life and abstract concepts.

It is shown that the proposed methodology represents a new toolkit for comparing and interpreting vector models of words and texts for the Russian language. Models that have undergone the testing procedure are included in a rating with public results and a report on the degree of language proficiency, on the degree of the quality of proficiency in skills and the dependence of this quality on various phenomena of the language, which are presented in the examples for testing. The rating made it possible to build target benchmarks for the development of Russian-language vector models: over 2 years of existence, the average quality indicator for all skills in the tested models rose from 49.5% (for the BERT-DeepPavlov model, the best solution at the time of launch) to 75.5% (for the best current solution, Golden transformer models). Undoubtedly, this level of proficiency does not yet reach the same level among native speakers (81.1% on average), which is shown by mass testing; existing models also lag significantly behind in problems where a variety of language phenomena are used to complicate the formulation of the problem; they are also extremely sensitive to changes in problem formulations and the use of quantifiers, numerals, dative constructions, ellipsis, the presence of named entities - the presence of these non-obvious factors can significantly affect the quality of learning a skill with models (in particular, BERT-DeepPavlov). However, in machine reading tasks, where a more reliable result is achieved by using a larger amount of data in training, vector models show a result higher than the result of native speakers.

We believe that the methodology can be refined, supplemented and enriched with new types of tasks, annotation, and we welcome the continuation of work in this direction. The methodology makes it now possible to fill the existing methodological gap and move towards more reliable, interpretable practices of working with the "black box" of existing language models.

The compilation of potential new intellectual problems using theoretical knowledge about the language introduces us to the interdisciplinary, border zone between language modelling and intelligence modelling. Separability of the assessment of one from the other is possible with the formulation of new types of text benchmarks, with a controlled set of linguistic properties and decision conditions.

Currently, textual benchmarks form a multidisciplinary field that combines areas such as linguistics, machine learning, and philosophy. The so-called "summer of artificial intelligence", whose offensive is associated with the development and popularization of vector models of words and texts, requires from the named disciplines a new methodology for recording new achievements, including systems in Russian.

References

- Turney, P. D., P. Pantel. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37(1), 141-188. (2010)
- Mikolov, T., et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
- Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – C. 1532-1543.
- Conneau A. et al. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. arXiv preprint arXiv:1805.01070. – 2018.
- 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. – T. 8. – C. 842-866.
- Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham
- Kuratov, Y., Arkhipov, M. (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213.
- Wang A. et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding //arXiv preprint arXiv:1804.07461. – 2018.
- Wang A. et al. Superglue: A stickier benchmark for general-purpose language understanding systems //arXiv preprint arXiv:1905.00537. – 2019.
- Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – C. 5998-6008.
- Li Yi (2017) Avengers: Achieving Superhuman Performance for Question Answering on SQuAD 2.0 Using Multiple Data Augmentations, Randomized Mini-Batch Training and Architecture Ensembling. Stanford CS224N {Default} Project, Stanford University url
- Hassan H. et al. Achieving human parity on automatic Chinese to English news translation //arXiv preprint arXiv:1803.05567. – 2018.
- Xiong W. et al. Achieving human parity in conversational speech recognition //arXiv preprint arXiv:1610.05256. – 2016.
- Shavrina T. O. Methods of computational linguistics in the evaluation of artificial intelligence systems. Voprosy Jazykoznaniya, 2021. № 6. P.117-138.
- Shavrina T. Word vector models as an object of linguistic research. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019" Moscow, May 29–June 1, 2019
- Shavrina T., Fenogenova A., Emelyanov A., Shevelev D., Artemova E., Malykh V., Mikhailov V., Tikhonova M., Chertok A., Evlampiev A. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. P. 4717-4726.
- Shavrina T., Emelyanov A., Fenogenova A., Fomin V., Mikhailov V., Evlampiev A., Malykh V., Larin V., Natekin A., Vatulin A., Romov P., Anastasiev D., Zinov N., Chertok A. Humans Keep It One Hundred: an Overview of AI Journey, in: Proceedings of The 12th Language Resources and Evaluation Conference Vol. 12. European Language Resources Association (ELRA), 2020. P. 2276-2284.

- Fenogenova A., ShavrinaT., Kukushkin A., Tikhonova M., Emelyanov A., Malykh V., Mikhailov V., Shevelev D., Artemova E.. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models (2021) A Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021” Moscow, 2021.

Appendix:

Appendix 1.
Linguistic diagnostics of the Russian SuperGLUE.

Meaning of the categories, with examples

Entailment

= Natural Language Inference

In general, we regard the NLI problem as one of judging what a typical human reader would conclude to be true upon reading the premise, absent the effects of pragmatics. Inevitably there will be many cases that are not purely, literally implied, but we want to build systems that will be able to draw the same conclusions as humans. Especially in the case of commonsense reasoning, which often relies on defeasible inference, this will be the case. We try to exclude particularly questionable cases from the data, and we do not use any sentences that are ungrammatical or semantically incoherent. In general, we use the standards set in the RTE Challenges and follow the guidelines of MultiNLI.

Given two sentences (a premise and hypothesis), we label them with one of two entailment relations:

Entailment: the hypothesis states something that is definitely correct about the situation or event in the premise.

Not Entailment: the hypothesis states something that might be correct or is definitely incorrect about the situation or event in the premise.
These definitions are essentially the same as what was provided to the crowdsourced annotators of the MultiNLI dataset. They rely on an assumption that the two sentences describe the same situation. However, a "situation" may involve multiple participants and actions, and the granularity at which we ask the described situations to be the same is somewhat subjective. The remainder of

this section describes the decisions we made when constructing the diagnostic dataset to decide these issues.

English Entailment:

- Writing Java is not too different from programming with handcuffs.
- Writing Java is similar to programming with handcuffs.

Russian Entailment:

- Написание кода на Java не слишком отличается от программирования в наручниках.
- Написание кода на Java подобно программированию в наручниках.

English - Not Entailment:

- When you've got snow, it's really hard to learn a snow sport so we looked at all the different ways I could mimic being on snow without actually being on snow.
- When you've got no snow, it's really hard to learn a snow sport so we looked at all the different ways I could mimic being on snow without actually being on snow.

Russian - Not Entailment:

- Когда у вас есть снег, очень сложно обучиться зимним видам спорта, так что мы искали все способы изучить без снега то, что я мог бы потом повторить на снегу.
- Когда у вас нет снега, очень сложно обучиться зимним видам спорта, так что мы искали все способы изучить без снега то, что я мог бы потом повторить на снегу.

Lexical semantics

These phenomena centre on aspects of word meaning.

Lexical Entailment

Entailment can be applied not only on the sentence level but the word level. For example, we say dog lexically entails animal because anything that is a dog is also an animal, and dog lexically contradicts cat because it is impossible to be both at once. This applies to all kinds of words (nouns, adjectives, verbs, many prepositions, etc.) and the relationship between lexical and sentential entailment has been deeply explored, e.g., in systems of Natural Logic. This connection often hinges on monotonicity in language, so many Lexical Entailment examples will also be tagged with one of the Monotone categories, though we do not do this in every case (see Definite Descriptions and Monotonicity).

- Falcon Heavy is the smallest rocket since NASA's Saturn V booster, which was used for the Moon missions in the 1970s.

- Falcon Heavy is the largest rocket since NASA's Saturn V booster, which was used for the Moon missions in the 1970s.

Morphological Negation

This is a special case of lexical contradiction where one word is derived from the other: from affordable to unaffordable, agree to disagree, etc. We also include examples like ever and never. We also label these examples with Negation or Double Negation, since they can be viewed as involving a word-level logical negation.

- Brexit is a reversible decision, Sir Mike Rake, the chairman of WorldPay and ex-chairman of BT group, said as calls for a second EU referendum were sparked last week.
- Brexit is an irreversible decision, Sir Mike Rake, the chairman of WorldPay and ex-chairman of BT group, said as calls for a second EU referendum were sparked last week.

Factivity

Propositions appearing in a sentence may be in any entailment relation with the sentence as a whole, depending on the context in which they appear.

- All speech is political speech.
- Joan doubts that all speech is political speech.

In many cases, this is determined by lexical triggers (usually verbs or adverbs) in the sentence. For example,

- I recognize that X entails X
- I did not recognize that X entails X
- I believe that X does not entail X
- I am refusing to do X contradicts I am doing X
- I am not refusing to do X does not contradict I am doing X
- I almost finished X contradicts I finished X
- I barely finished X entails I finished X Constructions like the one with recognize are often called factive, since the entailment (of X above, regarded as a presupposition) persists even under negation. Constructions like the one with refusing above are often called implicative, and are sensitive to negation. There

are also cases where a sentence (non-)entails the existence of an entity mentioned in it, e.g.,

- "I have found a unicorn" entails "A unicorn exists"
- "I am looking for a unicorn" does not necessarily entail "A unicorn exists" Readings where the entity does not necessarily exist are often called intensional readings, since they seem to deal with the properties denoted by a description (its intension) rather than being reducible to the set of entities that match the description (its extension, which in cases of non-existence will be empty). We place all examples involving these phenomena under the label of Factivity. While it often depends on context to determine whether a nested proposition or existence of an entity is entailed by the overall statement, very often it relies heavily on lexical triggers, so we place the category under Lexical Semantics.

Symmetry/Collectivity

Some propositions denote symmetric relations, while others do not; e.g.,

- "John married Gary" entails "Gary married John"
- "John likes Gary" does not entail "Gary likes John"

For symmetric relations, they can often be rephrased by collecting both arguments into the subject:

- "John met Gary" entails "John and Gary met"

Whether a relation is symmetric, or admits collecting its arguments into the subject, is often determined by its head word (e.g., like, marry or meet), so we classify it under Lexical Semantics.

- Republican lawmakers will ask President Trump to use a controversial White House framework as the baseline for a coming Senate debate on immigration policy.
- President Trump will ask Republican lawmakers to use a controversial White House framework as the baseline for a coming Senate debate on immigration policy.

Redundancy

If a word can be removed from a sentence without changing its meaning, that means the meaning of the words was more or less adequately expressed by the sentence; so, identifying these cases reflects an understanding of both lexical and sentential semantics.

- Tom and Adam were whispering loudly in the theatre.
- Tom and Adam were whispering in the theatre.

Named Entities

Words often name entities that exist out in the world. There are many different kinds of understanding we might wish to understand about these names, including their compositional structure (for example, the Baltimore Police is the same as the Police of the City of Baltimore) or their real-world referents and acronym expansions (for example, SNL is Saturday Night Live). This category is closely related to World Knowledge, but focuses on the semantics of names as lexical items rather than background knowledge about their denoted entities.

- The sides came to an agreement after their meeting in Europe.
- The sides came to an agreement after their meeting in Stockholm.

Quantifiers

Logical quantification in natural language is often expressed through lexical triggers such as every, most, some, and no. While we reserve the categories in Quantification and Monotonicity for entailments involving operations on these quantifiers and their arguments, we choose to regard the interchangeability of quantifiers (e.g., in many cases most entails many) as a question of lexical semantics.

- We consider all context words as positive examples and sample many negatives at random from the dictionary.
- We consider some context words as positive examples and sample negatives at random from the dictionary.

Logic

Once you understand the structure of a sentence, there is often a baseline set of shallow conclusions you can draw using logical operators. There is a long tradition of modelling natural language semantics using the mathematical tools of logic. Indeed, the development of mathematical logic was initially by questions about natural language meaning, from Aristotelian syllogisms to Fregean symbols. The notion of entailment is also borrowed from mathematical logic. So it is no surprise that logic plays an important role in natural language inference.

Propositional Structure

Negation, Double Negation, Conjunction, Disjunction, Conditionals

All of the basic operations of propositional logic appear in natural language, and we tag them where they are relevant to our examples:

Negation: The cat sat on the mat *contradicts* The cat did not sit on the mat.

When you have got snow, it is really hard to learn a snow sport so we looked at all the different ways I could mimic being on snow without actually being on snow.

> When you have got no snow, it is really hard to learn a snow sport so we looked at all the different ways I could mimic being on snow without actually being on snow.

Double negation: The market is not impossible to navigate *entails* The market is possible to navigate.

- The market is about to get harder, but possible to navigate.
- The market is about to get harder, but not impossible to navigate

Conjunction: Temperature and snow consistency must be just right *entails* Temperature must be just right.

- The patient bears some responsibility for successful care.
- Both doctor and patient bear some responsibility for successful care.

Disjunction: Life is either a daring adventure or nothing at all *does not entail, but is entailed by,* Life is a daring adventure.

- He has a blind trust.
- Either he has a blind trust or he has a conflict of interest.

Conditionals: If both apply, they are essentially impossible does not entail They are essentially impossible. Conditionals are a little bit more complicated because their use in language does not always mirror their meaning in logic. For example, they may be used at a higher level than the at-issue assertion: If you think about it, it is the perfect reverse psychology tactic entails It is the perfect reverse psychology tactic

- Pedro does not have a donkey.
- If Pedro has a donkey, then he beats it.

Quantifications

Universal, Existential Quantifiers are often triggered by words such as all, some, many, and no. There is a rich body of work modelling their meaning in mathematical logic with generalized quantifiers. In these two categories, we focus on straightforward inferences from the natural language analogues of universal and existential quantification:

Universal: All parakeets have two wings *entails, but is not entailed by* My parakeet has two wings.

- No one has a set of principles to live by.
- Everyone has a set of principles to live by.

Existential: Some parakeets have two wings *does not entail, but is entailed by* My parakeet has two wings.

- No one knows how turtles reproduce.
- Susan knows how turtles reproduce.

Monotonicity

Upward Monotone, Downward Monotone, Non-Monotone

Monotonicity is a property of argument positions in certain logical systems. In general, it gives a way of deriving entailment relations between expressions that differ on only one subexpression. In language, it can explain how some entailments propagate through logical operators and quantifiers. For example, note that pet entails pet squirrel, which further entails happy pet squirrel. We can demonstrate how the quantifiers a, no and exactly one differ with respect to monotonicity:

- "I have a pet squirrel" entails "I have a pet", but not "I have a happy pet squirrel".
- "I have no pet squirrels" does not entail "I have no pets", but does entail "I have no happy pet squirrels".
- "I have exactly one pet squirrel" entails neither "I have exactly one pet" nor "I have exactly one happy pet squirrel".

In all of these examples, the pet squirrel appears in what we call the restrictor position of the quantifier. We say: a is upward monotone in its restrictor: an entailment in the restrictor yields an entailment of the whole statement. no is downward monotone in its restrictor: an entailment in the restrictor yields an entailment of the whole statement in the opposite direction. exactly one is non-monotone in its restrictor: entailments in the restrictor do not yield entailments of the whole statement.

In this way, entailments between sentences that are built off of entailments of sub-phrases almost always rely on monotonicity judgments; see, for example, Lexical Entailment. However, because this is such a general class of sentence pairs, to keep the Logic category meaningful we do not always tag these examples with monotonicity; see Definite Descriptions and Monotonicity for details. To draw an analogy, these types of monotonicity are closely related to covariance, contravariance, and invariance of type arguments in programming languages with subtyping.

Richer Logical Structure:

Intervals/Numbers, Temporal There are some higher-level facets of reasoning that have been traditionally modelled using logic; these include actual mathematical reasoning (entailments based on numbers) and temporal reasoning (which is often modelled as reasoning about a mathematical timeline).

Intervals/Numbers: I have had more than 2 drinks tonight entails I have had more than 1 drink tonight.

- I failed my resolutions in 1995.
- I have failed my resolutions every year since 1997, and it is now 2008.

Temporal: Mary left before John entered *entails* John entered after Mary left.

- John entered after Mary left.
- Mary left before John entered.

Predicate-argument structure

An important component of understanding the meaning of a sentence is understanding how its parts are composed together into a whole. In this category, we address issues across that spectrum, from syntactic ambiguity to semantic roles and coreference.

Syntactic Ambiguity: Relative Clauses, Coordination Scope

These two categories deal purely with resolving syntactic ambiguity. Relative clauses and coordination scope are both sources of a great amount of ambiguity in English.

- Mao was chairman of the Communist Party from before its accession to power in 1949 until his death in 1976.
- The move marks an end to a system put in place by Deng Xiaoping in the 1980s to prevent the rise of another Mao, who was chairman of the Communist Party from before its accession to power in 1949 until his death in 1976.

Prepositional phrases

Prepositional phrase attachment is a particularly difficult problem that syntactic parsers in NLP systems continue to struggle with. We view it as a problem both of syntax and semantics since prepositional phrases can express a wide variety of semantic roles and often semantically apply beyond their direct syntactic attachment.

- On Sunday, Jane had a party.
- Jane had a party on Sunday.

Core Arguments

Verbs select for particular arguments, especially as their subject and object, which might be interchangeable depending on the context or the surface form. One example is the ergative alternation:

- "Jake broke the vase" entails "the vase broke".
- "Jake broke the vase" does not entail "Jake broke".

Other rearrangements of core arguments, such as those seen in Symmetry/Collectivity, also fall under the Core Arguments label.

Alternations: Active/Passive, Genitives/Partitives, Nominalization, Datives

All four of these categories correspond to syntactic alternations that are known to follow specific patterns in English:

- Active/Passive: I saw him is equivalent to He was seen by me and entails He was seen.
- Genitives/Partitives: the elephants foot is the same thing as the foot of the elephant.

- Nominalization: I caused him to submit his resignation entails I caused the submission of his resignation.v
- Datives: I baked him a cake entails I baked a cake for him and I baked a cake but not I baked him.

Ellipsis/Implicits

Often, the argument of a verb or other predicate is omitted (elided) in the text, with the reader filling in the gap. We can construct entailment examples by explicitly filling in the gap with the correct or incorrect referents. For example:

- Premise: Putin is so entrenched within Russia’s ruling system that many of its members can imagine no other leader.
- Entails: Putin is so entrenched within Russia’s ruling system that many of its members can imagine no other leader than Putin.
- Contradicts: Putin is so entrenched within Russias ruling system that many of its members can imagine no other leader than themselves. This is often regarded as a special case of anaphora, but we decided to split out these cases from explicit anaphora, which is often also regarded as a case of coreference (and attempted to some degree in modern coreference resolution systems).

Anaphora/Coreference

Coreference refers to when multiple expressions refer to the same entity or event. It is closely related to Anaphora, where the meaning of an expression depends on another (antecedent) expression in context. These phenomena have significant overlap, for example, with pronouns (she, we, it), which are anaphors that are co-referent with their antecedents. However, they also may occur independently, for example, coreference between two definite noun phrases (e.g., Theresa May and the British Prime Minister) that refer to the same entity, or anaphora from a word like other which requires an antecedent to distinguish something from. In this category we only include cases where there is an explicit phrase (anaphoric or not) that is co-referent with an antecedent or other phrase. We construct examples for these in much the same way as for Ellipsis/Implicits.

- George fell into the water.

- George went to the lake to catch a fish, but he fell into the water.

Intersectivity

Many modifiers, especially adjectives, allow non-intersective uses, which affect their entailment behaviour. For example:

- Intersective: He is a violinist and an old surgeon entails He is an old violinist and He is a surgeon
- Non-intersective: He is a violinist and a skilled surgeon does not entail He is a skilled violinist
- Non-intersective: He is a fake surgeon does not entail He is a surgeon

  Generally, an intersective use of a modifier, like old in old men, is one which may be interpreted as referring to the set of entities with both properties (they are old and they are men). Linguists often formalize this using set intersection, hence the name. It is related to Factivity; for example, fake may be regarded as a counter-implicative modifier, and these examples will be labelled as such. However, we choose to categorize intersectivity under predicate-argument structure rather than lexical semantics, because generally the same word will admit both intersective and non-intersective uses, so it may be regarded as an ambiguity of argument structure.

Restrictivity

Restrictivity is most often used to refer to a property of uses of noun modifiers; in particular, restrictive use of a modifier is one that serves to identify the entity or entities being described, whereas a non-restrictive use adds extra details to the identified entity. The distinction can often be highlighted by entailments:

- Restrictive: I finished all of my homework due today does not entail I finished all of my homework
- Non-restrictive: I got rid of all those pesky bedbugs entails I got rid of all those bedbugs.

Modifiers that are commonly used non-restrictively are appositives, relative clauses starting with which or who (although these can be restrictive, despite what your English teacher might tell you), and expletives (e.g. pesky). However, non-restrictive uses can appear in many forms.

Ambiguity in restrictivity is often employed in certain kinds of jokes (warning: language).

Knowledge

Strictly speaking, world knowledge and common sense are required on every level of language understanding, for disambiguating word senses, syntactic structures, anaphora, and more. So our entire suite (and any test of entailment) does test these features to some degree. However, in these categories, we gather examples where the entailment rests not only on correct disambiguation of the sentences but also application of extra knowledge, whether it is concrete knowledge about world affairs or more common-sense knowledge about word meanings or social or physical dynamics.

World Knowledge

In this category we focus on knowledge that can clearly be expressed as facts, as well as broader and less common geographical, legal, political, technical, or cultural knowledge. Examples:

- "This is the most oniony article I have seen on the entire internet" entails "This article reads like satire".
- "The reaction was strongly exothermic" entails "The reaction media got very hot".
- "There are amazing hikes around Mt. Fuji" entails "There are amazing hikes in Japan" but not "There are amazing hikes in Nepal".

Common Sense

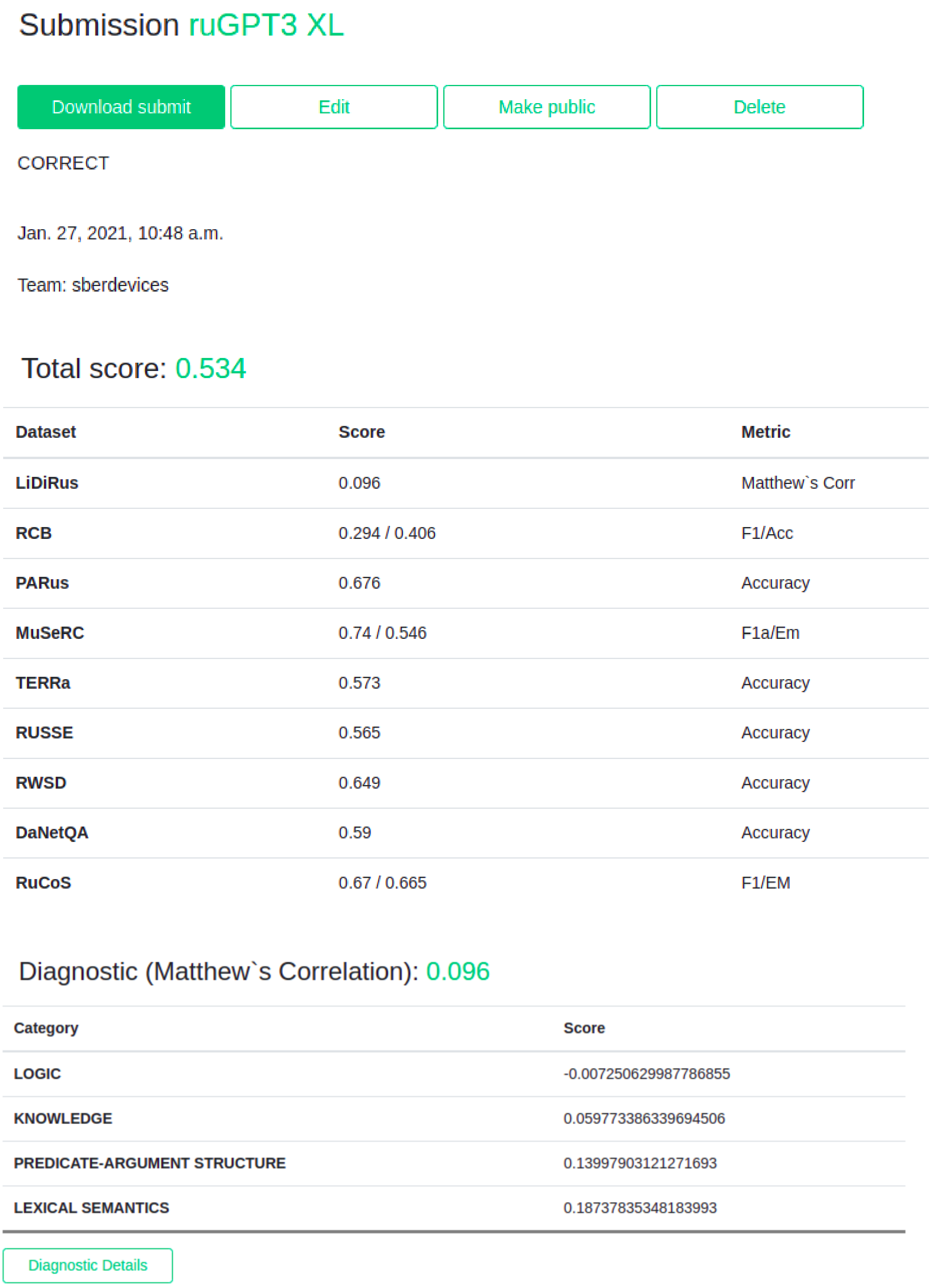
In this category we focus on knowledge that is more difficult to express as facts and that we expect to be possessed by most people independent of cultural or educational background. This includes a basic understanding of physical and social dynamics as well as lexical meaning (beyond simple lexical entailment or logical relations). Examples:

- "The announcement of Tillersons departure sent shock waves across the globe" contradicts "People across the globe were prepared for Tillersons departure".

- "Marc Sims has been seeing his barber once a week, for several years" entails "Marc Sims has been getting his hair cut once a week, for several years".
- "Hummingbirds are really attracted to bright orange and red (hence why the feeders are usually these colours)" entails "The feeders are usually coloured so as to attract hummingbirds".

Appendix 2

RSG submit procedure:



Submission ruGPT3 XL

CORRECT

Jan. 27, 2021, 10:48 a.m.

Team: sberdevices

Total score: 0.534

Dataset	Score	Metric
LiDiRus	0.096	Matthew`s Corr
RCB	0.294 / 0.406	F1/Acc
PARus	0.676	Accuracy
MuSeRC	0.74 / 0.546	F1a/Em
TERRa	0.573	Accuracy
RUSSE	0.565	Accuracy
RWSD	0.649	Accuracy
DaNetQA	0.59	Accuracy
RuCoS	0.67 / 0.665	F1/EM

Diagnostic (Matthew`s Correlation): 0.096

Category	Score
LOGIC	-0.007250629987786855
KNOWLEDGE	0.059773386339694506
PREDICATE-ARGUMENT STRUCTURE	0.13997903121271693
LEXICAL SEMANTICS	0.18737835348183993