Disambiguation of Russian Homographs with Transformers

Ivan Stolyarov and Olga Mitrofanova[0000-0002-3008-5514]

Saint-Petersburg State University, Saint-Petersburg, Russia

stolyarov.ivan.99@mail.ru, o.mitrofanova@spbu.ru

Abstract. The purpose of this work was to test transformer-based models BERT for homograph disambiguation in Russian, a long-standing issue in Text-To-Speech systems. The paper presents different types of Russian homographs and gives thorough analysis of existing methods of their disambiguation. A dataset of contexts from the Russian National Corpus for 28 homograph pairs was created and manually annotated. Three BERT models for the Russian language were selected and tested in two experiments. The results have shown that these models could achieve and outperform SOTA results in disambiguating homographs of all types on a relatively small training dataset. The pretrained models could also be used to disambiguate new pairs of intaparadigmatic homographs, absent in the original dataset.

Keywords: Russian Homographs, Homograph Disambiguation, Text-To-Speech, BERT.

1 Introduction

Speech synthesis is among the most rapidly developing areas of speech technologies and natural language processing (NLP). Extension of NLP technologies in conversational artificial intelligence increases the importance of text-to-speech (TTS) models and algorithms. TTS refers to conversion of orthographic text into audible speech. It is a multi-stage process, which includes linguistic processing and transcription of the incoming text which is required for further speech generation. The tasks of this stage include segmenting sentences, deciphering special symbols (*№*, *%*) and contractions (*т.д.* 'etc.', *км/ч* 'km/h'), homograph detection and disambiguation.

Homographs are properly defined as words of one or different parts of speech, which are the same in spelling, but differ in pronunciation and have different meanings, for example, *за́мок* 'castle'– *замо́к* 'lock', *му́ка* 'pain'– *мука́* 'flour', etc. [1]. We consider homography as a manifestation of ambiguity being the immanent property of natural language. In search of a reliable solution for homograph disambiguation in Russian we should take into account lexical-semantic ambiguity resolution, morphological and syntactic disambiguation, cf. [2-7].

Homograph disambiguation in synthesis systems involves selecting one of the possible options for the homograph and generating an appropriate transcription. Obviously, misreading the homograph (e.g., *за́мок* 'castle' instead of *замо́к* 'lock' in the sentence

Замок отнесли в мастерскую ‘The lock was taken to the workshop’) makes it difficult to perceive the synthesized speech and thus impairs the quality of synthesis.

Our study is aimed at improvement of homograph disambiguation quality in Russian. To achieve the goal, we propose an experimental dataset including contexts for Russian homographs, provide a detailed description of major classes of homographs and introduce a novel approach for homograph disambiguation based on contextualized BERT embeddings.

2 Homographs in Russian

In Russian, homography is associated with different stress in words that are spelled the same way, e.g. *а́тлас* ‘atlas’– *атла́с* ‘satin’, *ру́ки* ‘hands’ (pl, nom/acc) – *руки́* ‘hand’ (sg, gen), as well as with the practice of writing the letter "*е*" in place of the letter "*ё*", e.g. *все* ‘everybody’– *всё* ‘all’, *мел* ‘chalk’– *мёл* ‘swept’ (past, sg, masc). In some cases, both of these factors are involved: *бе́рег* ‘(river)bank’ – *берёг* ‘protected’ (past, sg, masc), *жены́* ‘wife’ (sg, gen) – *жёны* ‘wives’ (pl, nom).

According to lexical and grammatical correlation in the homograph pairs, the following types of homographs can be distinguished [7-9]:

- **Lexical (morphosyntactically congruent) homographs** – pairs of words that belong to separate lexemes and whose morphosyntactic values are identical, e.g. *за́мок* ‘castle’ – *замо́к* ‘lock’, *брони́ровать* ‘to reserve’– *бронирова́ть* ‘to armour’.
- **Intaparadigmatic** homographs – homographic wordforms belonging to the same lexeme, e.g. *жены́* ‘wife’ *(sg, gen)* – *жёны* ‘wives’ (pl, nom), *ва́рите* ‘(you) cook’ (ind, pres, 2 pers, pl) – *вари́те* ‘cook!’ (imper, pl).
- **Mixed (morphosyntactically incongruent) homographs** – homographs that belong to separate lexemes and whose morphosyntactic values are different. Homographs of this type can be divided into two categories. The first category includes pairs of words belonging to the same part of speech and differing in grammatical categories, e.g. *у́ха* ‘ear’ (masc, sg, gen) – *уха́* ‘fish soup’ (fem, sg, nom). The homographs in the second category belong to different parts of speech (e.g. a verb and a noun), which automatically results in a difference in grammatical meaning: *по́том* ‘sweat’ (noun, sg, instr) – *пото́м* ‘then’ (adv), *бе́рег* ‘(river)bank’ (noun, sg, nom/acc) – *берёг* ‘protected’ (verb, past, sg, masc).

Features representing different types of homographs are summarized in Table 1.

Table 1. Types of homographs and their features

Type of homographs	Different lexical meaning	Different morpho-syntactic values
Lexical	+	–
Intaparadigmatic	–	+
Mixed	+	+

3 Methods for Homograph Disambiguation

Various methods for homograph disambiguation have been proposed in theoretical research focused on homographs and in actual speech synthesis systems. All approaches agree that correct interpretation of the context surrounding a homograph is crucial but there are different ways of treating this issue. For example, rule-based methods imply manually compiling a set of rules that allow one to choose the correct way of reading a homograph depending on context neighbours. This approach was used, in particular, by B. Lobanov and colleagues in Multifon system [10]. There are methods based on contextual analysis performed on syntactic and semantic levels, or complex analysis of the nearest environment of a homograph or of the whole sentence. Thus, the research carried out by O. Khomitsevich and colleagues describes a similar mechanism of homograph disambiguation in the VitalVoice synthesis system developed by the Speech Technology Centre [11]. Finally, machine learning techniques can also be applied to disambiguate homographs. As an example, the model proposed in the work of K. Gorman and colleagues can be referred to in this respect: for each pair of homographs, a multinomial classifier was created, to which a set of features was fed (word context features, POS tags, capitalization feature) [12].

In 2021, M. Nikolis and V. Klimkov proposed the use of contextual word embeddings, which are produced by transformer models of BERT family (Bidirectional Encoder Representations from Transformers) [13]. Word embeddings from BERT are inherently context-sensitive: they encode information about lexical meaning and morphosyntactic values of the target word depending on its context. M. Nikolis and V. Klimkov tested two BERT models for English language and achieved state-of-the-art (SOTA) accuracy of 99.1% on the English homographs.

This paper presents the results of our attempt to adapt the experiment by M. Nikolis and V. Klimkov to disambiguate Russian homographs.

4 Dataset for Homograph Disambiguation

As far as we know, there is no specialized Russian dataset for training homograph disambiguation models, because homographs are usually considered together with homonyms in WSI/WSD tasks (cf. datasets for Dialogue Evaluation, especially RUSSE2018 [14]). Thus, we decided to collect and manually annotate data to fill in the existing gap. We used contexts from general fiction and newspaper articles presented in the Russian National Corpus [15]. Each context is represented by one sentence containing the target word (homograph). Our dataset includes the following groups of homographs:

- **Lexical homographs** – 4 homograph pairs: *а́тлас* 'atlas'– *атла́с* 'satin', *за́мок* 'castle'– *замо́к* 'lock', *му́ка* 'pain'– *мука́* 'flour, *хло́пок* 'cotton' – *хлопо́к* 'clap'. For each pair, we selected 100 contexts (50 corresponding to one homograph and 50 to another), in which homographs are represented in different wordforms: *за́мок – замо́к* (sg, nom/acc), *за́мка – замка́* (sg, gen), …, *за́мки – замки́* (pl, nom/acc), …
- **Intaparadigmatic homographs** – 3 subgroups, each of which included

4 homograph pairs belonging to the same accent type (see Table 2). 30 contexts were selected for each pair, so each subgroup consisted of 120 contexts in total. This is the only case in which we use as labels not the pronunciation variants of a homograph but the set of grammatical categories (sg gen vs pl nom/acc) conveyed by different accents in the wordforms (see Table 2). Different types of correlations among intaparadigmatic homographs are described, for example, by J. Kaliszan [16].

Table 2. Subgroups of intaparadigmatic homographs

I subgroup (masc) sg gen vs. pl nom/acc	**II subgroup (fem)** sg gen vs. pl nom/acc	**III subgroup (neut)** sg gen vs. pl nom/acc
ве́чера ‘evening’ – *вечера́* ‘evenings’	*головы́* ‘head’ – *го́ловы* ‘heads’	*ме́ста* ‘place’ – *места́* ‘places’
го́рода ‘city’ – *города́* ‘cities’	*горы́* ‘mountain’ – *го́ры* ‘mountains’	*мо́ря* ‘sea’ – *моря́* ‘seas’
о́строва ‘island’ – *острова́* ‘islands’	*руки́* ‘hand’ – *ру́ки* ‘hands’	*по́ля* ‘field’ – *поля́* ‘fields’
по́езда ‘train’ – *поезда́* ‘trains’	*страны́* ‘country’ – *стра́ны* ‘countries’	*сло́ва* ‘word’ – *слова́* ‘words’

- **Mixed homographs (nouns)** – 4 pairs of nouns: *бе́лка* ‘squirrel’ (sg, nom)– *белка́* ‘protein’ (sg, gen), *ви́ски* ‘whiskey’ (sg, nom) – *виски́* ‘temples’ (pl, nom/acc), *го́ре* ‘sorrow’ (sg, nom/acc) – *горе́* ‘mountain’ (sg, loc), *ду́ша* ‘shower’ (sg, gen) – *душа́* ‘soul’ (sg, nom). For each pair, we collected 100 contexts.
- **Mixed homographs (verbs)** – 4 pairs of verbs: *вы́купать* ‘to bathe’ – *выкупа́ть* ‘to ransom’, *вы́читать* ‘find by reading’ – *вычита́ть* ‘to substract’, *разре́зать* ‘to cut up’ (perf) – *разреза́ть* ‘to cut up’ (imperf), *сбе́гать* ‘to run for smth’ – *сбега́ть* ‘to run away’. For each pair, we selected 100 contexts presenting different conjugated forms of these verbs (*вы́купаю* – *выкупа́ю*, *вы́купаешь* – *выкупа́ешь*, …, *вы́купал* – *выкупа́л*, …).
- **Mixed homographs (differents POS)** – 4 pairs of words belonging to different parts of speech: *бе́рег* ‘(river)bank’ (noun, sg, nom/acc) – *берёг* ‘protected’ (verb, past, sg, masc), *ве́сти* ‘news’ (noun, pl, nom/acc) – *вести́* ‘to lead’ (verb, inf), *зна́ком* ‘sign’ (noun, sg, instr) – *знако́м* ‘familiar’ (adj, sg, masc), *по́том* ‘sweat’ (noun, sg, instr) – *пото́м* ‘then’ (adv). 100 contexts were selected for each pair.

To sum up, our dataset includes 1960 contexts for 28 homograph pairs of different types.

In addition, in the Experiment II we also tested 20 new contexts for each subgroup of intaparadigmatic homographs. These contexts included pairs of homographs that belonged to the same grammatical correlation (and the same accent type) but was not present in the data on which the model was trained (cf. Table 3).

Table 3. New pairs of homographs for the Experiment II

Subgroup of intaparadigmatic homographs	New pair of homographs
I (masc.)	*а́дреса* 'address' (sg, gen) – *адреса́* 'addresses' (pl, nom/acc)
II (fem.)	*игры́* 'game' (sg, gen) – *и́гры* 'games' (pl, nom/acc)
III (neutr.)	*се́рдца* 'heart' (sg, gen) – *сердца́* 'hearts' (pl, nom/acc)

5 Baseline Results of Homograph Disambiguation

First of all, we decided to test existing solutions for homograph disambiguation on our dataset. The following models were selected for disambiguating different groups of homographs:

- **RuWordNet thesaurus** for lexical homographs (group 1). RuWordNet is a thesaurus of the Russian language, which contains synsets (sets of synonyms) for nouns, verbs and adjectives and establishes different semantic relations between synsets of the same part of speech (hyponym-hypernym, instance-class, etc.) [17]. Therefore, RuWordNet can be used to disambiguate lexical homonyms and homographs [18].
- **RNN Morph** for homographs from groups 2, 3, 5. RNN Morph is a morphological analyzer (POS tagger) for Russian and English languages, which is based on neural networks and dictionary-lookup systems (pymorphy2, NLTK) and showed the best results on the MorphoRuEval-2017 competition [18]. We used the RNN Morph to disambiguate homographs that differ in morphosyntactic categories and are expected to be tagged differently.
- **SpaCy** for mixed verbal homographs from the group 4. SpaCy is an open-source software library for NLP. We used a Russian language model from spaCy (*ru_core_news_md*) which tagged verbs from group 4 with their aspect (perfective/imperfective) as it is a grammatical category that makes a difference between them [20].

The results of these models on different homograph group from our dataset are shown in Table 4. As it can be seen from the table, baseline models fail to achieve acceptable results and, from our observations, often give preference simply to the most frequent option, e.g. *а́тлас* 'atlas' instead of *атла́с* 'satin' (in case of RuWordNet); or the tag 'verb' instead of 'noun' for *ве́сти* 'news'– *вести́* 'to lead' (in case of RNN Morph), etc.

Table 4. Baseline accuracies

Group of homographs	RuWordNet	RNN Morph	SpaCy
1. Lexical	62%	–	–
2. Intaparadigmatic	–	78%	–
3. Mixed (nouns)	–	78%	–
4. Mixed (verbs)	–	–	49%
5. Mixed (different POS)	–	81%	–

6 Model Description

In our experiments, we used contextualized embeddings produced by the following pre-trained BERT models for Russian:

- **RuBert-base** (12-layer, 768-hidden, 12-heads, 180M parameters) from the library DeepPavlov represents the target word as a vector of 768 dimensions [21]. The model was trained on Russian Wikipedia and news data.
- **RuBert-base finetune** (12-layer, 768-hidden, 12-heads, 180M parameters) also produces word embeddings of 768 dimensions [22]. This model was trained by SberDevices team.
- **RuBert-large finetune** (24-layer, 1024-hidden, 16-heads, 430M parameters) generates vectors of greater dimensions (1024) [23]. This model was also trained by SberDevices team.

We followed the same experimental scheme while working with the given three models. After the dataset was loaded, each sentence with a homograph was tokenized and special tokens were added to it: [CLS] at the beginning and [SEP] at the end. After the tokenized sentence was processed by a BERT model, an embedding corresponding to the token-homograph was extracted from the last (12^{th} or 24^{th}) layer. It was the way we formed a new dataset: the objects (X) were vectors of 768 dimensions (768 features) and the target variables (Y) were pronunciation variants (as mentioned above, in the case of intaparadigmatic homographs we used a set of grammatical categories as target variables). The resulting dataset was split into training and test data with the ratio of 4:1. We used logistic regression with L2 regularization for embedding classification task. We evaluated the model by calculating accuracy (percentage of correctly classified examples). Accuracy is calculated by the formula:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} * 100\%,$$

where *TP* is the number of true positives, i.e., correct predictions for the class I; *TN* stands for the number of true negatives, i.e., correct predictions for the class II; *FP* and *FN*, respectively, the number of false positives and false negatives (wrong predictions).

7 Experiments and Results

7.1 Experiment I. Disambiguating Homographs of Different Types

Experiments have shown that BERT models could achieve at least 95% of accuracy in disambiguating different groups of homographs. The best average result (97,6%) was obtained by the model RuBert-large finetune, but it could be seen that some groups of homographs (1-3) were better handled by the model RuBert-base finetune. It may also be noted that verbal homographs from the group 4 constituted the greatest difficulty for all three models (85%, 88%, 95% respectively).

Table 5. Results of the Experiment I

Type of homographs	rubert-base-cased	rubert-base (Sber)	rubert-base large
1. Lexical	98%	**99%**	**99%**
2. Intaparadigmatic	96%	**99%**	96%
3. Mixed (nouns)	98%	**100%**	99%
4. Mixed (verbs)	85%	88%	**95%**
5. Mixed (different POS)	98%	98%	**99%**
Average	95,0%	96,8%	**97,6%**

7.2 Experiment II. Disambiguation of Intaparadigmatic Homographs Absent in the Original Dataset

Each model trained on a subgroup of intaparadigmatic homographs was fed with new contexts including a pair of homographs representing the same grammatical correlation (and the same accent type) but absent in the original dataset. The results we obtained (cf. Table 6) indicate that the trained model is capable of predicting grammatical categories (in our case *sg, gen* vs. *pl, nom/acc*) for new homographs and, consequently, their accent type which makes it possible to disambiguate them.

Table 6. Results of the Experiment II

Subgroup of intaparadigmatic homographs	Accuracy on the test data (%)	Accuracy on the new data (%)
I (masc.)	94,0	**92,5**
II (fem.)	95,0	**100,0**
III (neutr.)	100,0	**97,5**

8 Discussion

The experiments show that the proposed approach could achieve and outperform SOTA results of homograph disambiguation. It was also demonstrated that a relatively small dataset (100 sentences) is sufficient to train a classifier for this task, provided that both pronunciation variants of a homograph are equally represented. Another advantage of our model is its universality: we use the same algorithm to disambiguate all types of homographs, regardless of their distinctive features (lexical and/or morphosyntactic), without any supplementary resources being necessary. Finally, we showed in Experiment II that the trained model could also disambiguate new pairs of intaparadigmatic homographs, absent in training data.

However, the proposed approach has some limitations. Since BERT encodes all kinds of contextual information (lexical, grammatical, etc.) in one embedding for the token of interest (homograph), we could not easily figure out from the results which contextual features played a crucial role in disambiguation or, on the contrary, which ones did not prove to be significant.

One may also notice that verbal homographs (group 4) still pose a challenge for a disambiguation system. This seems to be due to the specificity of the verb aspect and the necessity to analyze distant syntactic relations. For example, in the sentence *Она **разрезала** его [пласт капусты] ножиком на слоистые куски, доставала вилки, хлеб* ‘She cut it [some cabbage] into flaky pieces with a knife, pulled out forks and bread’, the homograph *разрезала* is an imperfective verb, thus the stress has to be put on the third syllable (*разреза́ла*). The information about verb aspect of *разрезала* could be extracted from the word *доставала*, also an imperfective verb, which shows that one puts the emphasis on the process of both actions rather than on the result (normally expressed by perfective aspect). However, our model seems not to take into account the word *доставала* because of the great distance between two verbs.

As described in the section 4, the context of each homograph was limited to one sentence, though one may fairly assume that this breadth of context is not sufficient in some cases. Indeed, we encountered a few sentences which did not contain enough lexical and/or grammatical features for homograph disambiguation. Let us examine the sentence *Пролетариат **замков** не признает.* ‘Proletarians do not accept any castles/locks’. It seems to us that both meanings (‘castles’ and ‘locks’) could be appropriate in this sentence. It’s the broader context that makes it possible to figure out that ‘locks’ (*замко́в*) is correct: *Заходи! Открыто всегда! Пролетариат* ***замков*** *не признает. Дверь заскрипела, и на пороге маленькой комнаты…* 'Come in! It's always open! Proletarians do not accept any locks. The door creaked open and on the threshold of the small room…'. The same issue is also one of the sources of difficulty in disambiguating verbal homographs. We suggest considering an example with the homograph pair *разре́зать – разреза́ть* that has been already mentioned: *Дед Исаак очень много ел. Батоны* ***разрезал*** *не поперек, а вдоль. В гостях бабка Рая постоянно за него краснела.* 'Grandpa Isaac ate a lot. He cut the loaves lengthwise instead of crosswise. Grandma Raya always blushed for him when they were on a visit'. The sentence with the target homograph (*Батоны* ***разрезал*** *не поперек, а вдоль*) might be interpreted in two ways: it could refer to the result of Isaac's action (perfective verb) or to Isaac's habit (imperfective verb). We could resolve the ambiguity by taking into account the neighbouring sentences: the imperfective verbs *ел* and *краснела* explicitly confirm the second interpretation, thus the option *разреза́ла* has to be chosen. These examples demonstrate that limiting the context to one sentence is not always the appropriate solution.

9 Conclusion and Future Work

In this study we used contextualized word embeddings produced by BERT models in disambiguation of Russian homographs. We showed that the proposed approach reaches the level of SOTA models (up to 97,6% of accuracy) and has a wide range of advantages. Yet we also discuss some of its limitations. It is anticipated that the developed approach could improve the quality of speech synthesis.

Directions for future work deal with:

- considering other groups and subgroups of Russian homographs;
- improving the obtained results by a more careful selection of contexts and/or increasing the dataset;
- analysis of challenging cases of disambiguation;
- analysis of how the phrase boundaries and context window size affects results of disambiguation;
- implementation of the developed algorithm in a speech synthesis system.

Acknowledgements. Research is performed with support of SPbSU Scientific Project № 75254082 «Modeling the communicative behavior of residents of a Russian metropolis in the socio-speech and pragmatic aspects with the use of artificial intelligence methods».

References

1. Kozhina, M.N.: Stylistic encyclopedic dictionary of the Russian language. Moscow. (2003)
2. Lyashevskaya, O., Afanasev, I., Rebrikov, S., Shishkina, Y., Suleimanova, E., Trofimov, I., Vlasova, N.: Disambiguation in context in the Russian National Corpus: 20 years

later. In: Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference «Dialogue 2023». RSUH, Moscow. P. 307-318. (2023)

3. Alexeyevsky, D., Temchenko, A.V.: WSD in Monolingual Dictionaries for Russian Word Net. In: Proceedings of the Eighth Global Word Net Conference. Bucharest, Romania. P. 9-14. (2016)
4. Bolshina, A., Loukachevitch, N.: All-words Word Sense Disambiguation for Russian Using Automatically Generated Text Collection. In: Cybernetics and Information Technologies. Vol. 20. Issue 4. P. 90-107. (2020)
5. Lopukhin, K.A., Lopukhina, A.A.: Word Sense Disambiguation For Russian Verbs Using Semantic Vectors And Dictionary Entries. In: Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference «Dialogue 2016». RSUH, Moscow. P. 393-404. (2016)
6. Panicheva, P.V., Protopopova, E.V., Mirzagitova, A.R., Mitrofanova, O.A.: Development of an NLP Toolkit for Morphological Analysis of Russian Text Corpora Based on PyMorphy and NLTK. In: Proceedings of the International Conference «Corpus Linguistics – 2015». Saint-Petersburg. P. 361-373. (2015)
7. Inshakova, Ye.S., Iomdin, L.L., Mityushin, L.G., Sizov, V.G., Frolova, T.I., Tsinman, L.L.: The Syntagrus Today. In: Proceedings of Vinogradov Russian Language Institute of the Russian Academy of Sciences. Issue 21. P. 14-40. (2019)
8. Kolesnikov, N.P.: The system of word homonyms in Russian. In: 12th session of the Faculty of Philology. Work plan and abstracts. Tbilisi University Press, Tbilisi. P. 27-29. (1968)
9. Melnikova, A.I.: On the subject of Russian homographs. In: Russian language at school. Issue 4. P. 80-84. (1974)
10. Tsirulnik, L.I., Barbuk, S.G., Lobanov, B.M.: Statistical analysis and contextual rules of disambiguation of graphic homonymy in TTS. In: Computational Linguistics and Intellectual Technologies: Paper from the Annual International Conference «Dialogue 2009». RSUH, Moscow. P. 530-536. (2009)
11. Khomitsevich, O.G., Rybin, S.V., Anichkin, I.M.: Using linguistic analysis to normalize text and disambiguate homonyms in the Russian speech synthesis system. In: Proceedings of Higher Education Institutions. Instrumentation engineering. Vol. LVI. Issue 2. P. 42-46. (2013)
12. Gorman, K., Mazovetskiy, G., Nikolaev, V.: Improving homograph disambiguation with supervised machine learning. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan. European Language Resources Association (ELRA). P. 1349-1352. (2018)
13. Nicolis, M., Klimkov, V.: Homograph disambiguation with contextual word embeddings for TTS systems. In: Proceedings of the 11th ISCA Speech Synthesis Workshop (SSW 11). P. 222-226. (2021)
14. URL: https://www.dialog-21.ru/evaluation/2018/disambiguation/
15. URL: https://ruscorpora.ru/
16. Kaliszan, J.: Grammatical substantive homography in Russian. In: Studia Rossica Posnaniensia. Vol. 38. P. 119-123. (2013)
17. https://ruwordnet.ru/en
18. https://github.com/avidale/python-ruwordnet
19. https://github.com/IlyaGusev/rnnmorph
20. https://spacy.io/models/ru
21. https://huggingface.co/DeepPavlov/rubert-base-cased
22. https://huggingface.co/ai-forever/ruBert-base
23. https://huggingface.co/ai-forever/ruBert-large