








Developing a Question Answering System on the Material of Holocaust Survivors' Testimonies in Russian

Liudmila Bukreeva , Daria Guseva , Mikhail Dolgushin  ^(✉),
Vera Evdokimova , and Vasilisa Obotnina 

Saint Petersburg State University, Universitetskaya emb., 7-9, St. Petersburg 199034, Russia
v.evdokimova@spbu.ru

Abstract. The paper makes use of the annotated task-oriented corpus of Holocaust testimonies in Russian (ruOHQA) to train a question-answer neural network model. We start from data preprocessing, present statistical analysis of the collected corpus for approximately 1500 pairs of questions and answers and describe its strengths and limitations. Also, we carry out experiments on automatic processing of the ruOHQA corpus using pre-trained transformer-based neural network models. Finally, we explore the capability of several models to generate simplified high-quality answers to questions and compare their results. The kind of research we present allows us to extract knowledge from oral history archives more productively.

Keywords: Question Answering · Corpora · Visual History Archives

1 Introduction

Question Answering Systems (QAS) have become an important field of research in natural language processing combining such tasks as information extraction and machine learning. QASs help to obtain answers to questions of interest asked in natural language which may be essential for various specific research issues [1]. In our paper we concentrate on extracting answers to questions from oral history archives. Oral history preserves historical records [2] in the form of an interview with people who witnessed historically significant events.

Oral history data encompasses multiple topics, one of which is Holocaust testimonies. The large amount of data includes The Visual History Archive of Holocaust testimonies compiled by the USC Shoah Foundation [3] with over 7,000 multimedia recordings and 25 freely available interviews with Holocaust survivors from the Yad Vashem Foundation in Russian [4]. Our aim is to summarize facts and stories from the interviews provided by the Yad Vashem Foundation. Our choice was made due to the fact that most of the video interviews contain manually typed subtitles. We enable the analysis and interpretation of these oral history archives by collecting tagged corpus for the presented records. It helps to satisfy the stable interest in materials of such kind [5–7] by turning this large amount

of data into a more attainable form as well as affects the accuracy of the forthcoming QAS.

QASs perform at their best when they deal with structured knowledge bases [1]. Therefore, we gather the Question-Answer corpus ruOHQA and further use this dataset to train neural network models for the question answering task. The corpus contains over 1,500 automatically gathered entries, further manually aligned and labeled by experts.

2 Related Work

There are several efforts to build QASs able to retrieve information from visual history data. The research [15] develops the dialogue system based on the international project MALACH (Multilingual Access to Large Spoken Archives). The QAS for English and Czech parts of the MALACH archive of Holocaust testimonies allows one to obtain answers using spoken natural language queries.

The paper [6] presents the QA model formulating queries in a natural language. Due to the colloquial form of speech in the researched mMQA corpus with 8914 entries of questions and answers, the final accuracy turns out to be very limited. The experimental results indicate that the further research on building QASs for oral history data remains relevant.

We were inspired by the paper [15], thus, our motivation was to retrieve information from oral history archives of Holocaust testimonies in Russian. To the best of our knowledge, there are no similar QA datasets and QAS for Russian oral history archives of such kind mentioned in the literature.

The main dataset used to solve the QA problem for Russian is SberQuAD [16] with approximately 50,000 question-answer pairs, which are splitted on 45,3 k train, 5,04 validation and 23.9k test rows. This reading comprehension dataset contains Wikipedia articles and questions to its segments posed by a group of crowdworkers. Each question presupposes an answer from the corresponding reading passage, however, might remain unanswered. The methodology used to create SberQuAD was similar to what was used for the development of the English SQuAD corpus [17], and SQuAD 2.0 [18]. The structure of these datasets has shown the significance of including unanswerable questions in the corpora.

We will follow the practice of earlier works published. Our tasks involve creating the QA dataset and applying it to train the QAS.

3 Corpus Creation

Research on building QAS has always been constrained by the limited availability of structured training data. Thus, collecting appropriate textual data and structuring it was the first step required in our work.

3.1 Text Collection

We started our corpus creation from collecting video recordings of interviews with Holocaust survivors from the Yad Vashem Foundation [4]. We decided to add into our

corpus only recordings containing subtitles preprocessed by specialists of foundation. There were 4 recordings among 25 with automatic subtitles. We could not add them to the corpus since speech recognition technology is the error prone process, consequently, the quality of such subtitles might be low. As a result, we gathered 21 transcribed recordings with the total duration of approximately 26 h. We extracted all the subtitles from each recording. The total size of the unprocessed corpus reached 20200 unique pairs of questions and answers.

To identify video recordings and subtitle files, we assigned them an individual identification code. Keeping all the video materials in order was also necessary to further clarify the controversial points appearing during the corpus annotation. In particular, it ensured that potential context gaps, such as interruptions by the interviewer during the interviewee's response, were not overlooked.

3.2 The Annotation

The first step of dataset annotation involved dividing the interviewer's speech from one of the narrator's. The material in our unprocessed corpus already was in Russian and contained punctuation, which allowed us to conduct preliminary annotation by rules. To extract the context of the expected replies, we followed the basic assumption that an interrogative sentence might be followed by an answer. Thus, we created a new corpus entry in case there was a question mark in a previous sentence. As a matter of course, questions following one another and building an interviewer's speech turned out to be divided, thereby we encountered false answer selection: e.g. *“Как ее звали? Вы помните?”*—*“What was her name? Do you remember?”*. In this example *“Do you remember?”* was automatically extracted as the answer, although we clearly understand that it is the question. Along with this, we mentioned that the false detection of questions occurred as well. It happened when there were rhetorical questions or questions within the context of a story in narrator's answer speech: e.g. *“и ему говорят: “Слушай, ты этого мальчика знаешь?” Он говорит...”*—*“and they say to him: “Listen, do you know this boy?” He says...”*.

With help of initial automatic preprocessing we extracted 4228 pairs of question-answer contexts with the preservation of the indexes from the subtitles. As might be expected, many errors occurred due to the specifics of automatic preprocessing, which does not take into account the peculiarities of the spoken form of the interviews and peculiar coloring of the speeches caused by the age of the narrators. These features made us decide to annotate the corpus manually using an expert assessment.

The next step of our work required dividing the corpus into parts equal in number of entries further given to 4 experts to annotate manually. The experts had access to all the materials and were required to act according to a unified set of instructions. The manual annotation included the following tasks: correction of errors caused by the automatic preprocessing and construction of the specialized format for our corpus useful for forthcoming QA training purposes. In case question entries were inaccurately assigned to the context of the answer or were not punctuated, we created new question entries for them. Punctuation was maintained if necessary. The context of the answer was cleared of possible interviewer's remarks along with grammatical and orthographic mistakes.

As a result, we managed to annotate 1555 entries that were composed into the Russian Oral History Question Answering dataset (ruOHQA). An example of the record with translation into English is presented in Table 1. The same structure is followed in every entry of the corpus.

Table 1. Corpus sample with translation.

id	question	answer	context
279_297	Это уже в какое время года было?	Это было, уже я пошла в школу, это к сентябрю.	Это я тебе сейчас скажу... Это было, уже я пошла в школу, это к сентябрю. Мы все лето, мы все время убегали от немцев. Нас даже там не высадили...
279_297	What time of year was this?	It was, I already went to school, this was by September.	I'll tell you now... It was, I already went to school, this was by September. We spent the whole summer, we ran away from the Germans all the time, they didn't even drop us off...

A corpus entry consists of four columns. The first column has unique indexes of the interrogative speech from the subtitles. The second column contains the interviewer's question to the narrator. The third column includes only direct answers to the interviewer's question. Finally, the fourth column contains the detailed context of the answer provided by the narrator within their story during the interview.

4 Data Analysis

In Table 2 we compare the ruOHQA dataset to the similar Russian QA corpus SberQuAD. A comprehensive description of all possible SberQuAD features is given in [16]. We compare such parameters as the average question, answer and context length in both QA corpora. As can be seen from Table 2, the average question length turns out to be similar in both datasets, while the average answer length shows noticeable differences. We explain the longer average length of answers in the ruOHQA corpus by the fact that the narrators make their speech more extended and often less concise by going into the details.

Table 2. Statistics of the ruOHQA and SberQuAD datasets.

Dataset	Total number of samples	Avg. question length (words)	Avg. answer length (words)	Avg. context length (words)
ruOHQA	1,555	7,076	5,444	22,858
SberQuAD	50,364	8,613	2,433	98,666

In order to analyze the content of the ruOHQA corpus, we counted 30 most common tokens in its question and answer parts. We lemmatized words with the Python library `pymorphy2` [20] to conduct some preliminary processing. Further processing included removing of stop words, namely prepositions and conjunctions. In this way, we were able to extract only the tokens that were necessary to our query.

Finally, we obtained the token frequency graphs in Fig. 1 and Fig. 2 with `nlTK`, the natural language processing library in Python [21].

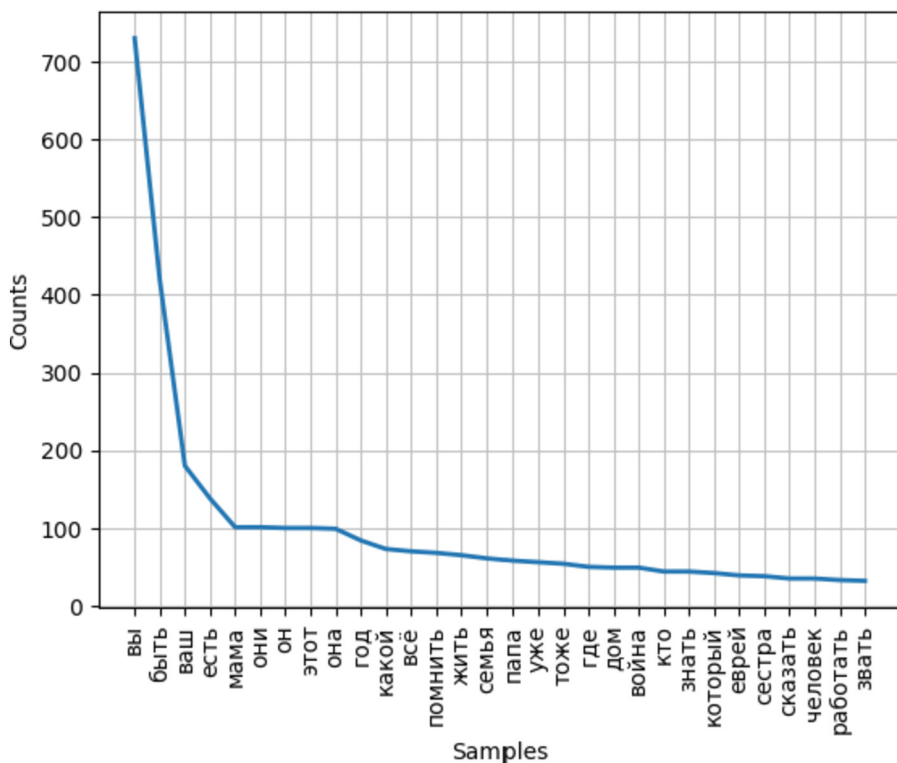
**Fig. 1.** Frequency distribution for 30 most common tokens in ruOHQA questions.

Figure 1 shows the most frequent lemmatized words counted in the question part of the ruOHQA corpus. We identify words connected with Holocaust, e.g. ‘война’ (war),

‘помнить’ (remember), ‘еврей’ (jew), with family or relatives, e.g. ‘мама’ (mom), ‘сестра’ (sister) etc. We intentionally did not remove pronouns and interrogative words in the questions’ frequency list before counting, as they may also express an interviewer’s appeal to narrators.

The solid curve in Fig. 2 represents the 30 most common lemmatized words found in answers of the ruOHQA corpus. We notice similar tokens including verbs related to memory, e.g. ‘знать’ (to know), ‘помнить’ (to remember), nouns naming family members, e.g. ‘мама’ (mom), ‘папа’ (dad), ‘бабушка’ (grandma) etc. An important frequently used word is ‘еврей’ (jew). It shows us the nationality of narrators and remains a core concept for specific topics discussed. Eventually, the frequency usage turns out to be quite representative for the content in our corpus based on the interviews with Jewish Holocaust survivors.

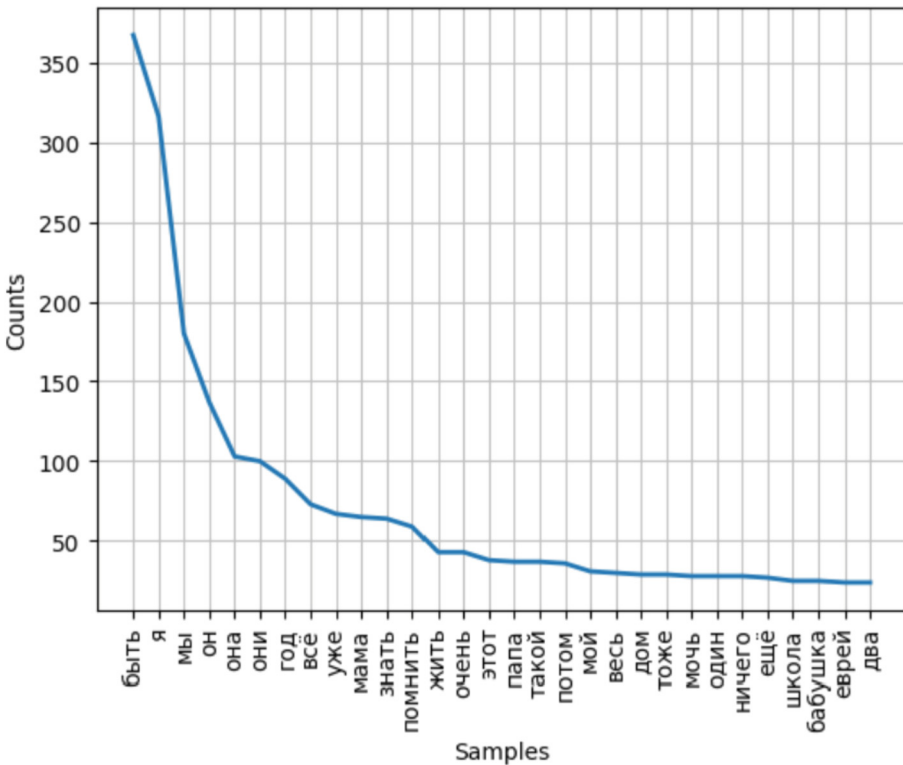


Fig. 2. Frequency distribution for 30 most common tokens in ruOHQA answers.

5 Experiments

Once we have the ruOHQA corpus ready for training our QAS, we start the experimental setup of the models. The initial step of experiments included fine-tuning using our corpus. We have selected the distilled versions of ruBERT models pretrained on informal texts

from DeepPavlov [22] in the Huggingface framework [23]. Our choice has been made based on several reasons: 1) the informal texts used for pretraining the models correspond to the dialogue structure of our ruOHQA dataset; 2) the distilled versions of the models perform with relatively quick learning rate; 3) the performance of the distilled versions keeps up with the full models [24]; 4) finally, the ruBERT models show the high level of accuracy in automatic language processing tasks in Russian [25].

The ruOHQA dataset was divided into train and test subsets in the ratio of 0.7 to 0.3, for training and evaluation of the resulting model respectively. Thus, we received 929 entries of the corpus for training the model and 399 entries for testing. The size of the ruOHQA dataset is limited for the model to be trained only on it. Hence, we decided to evaluate the accuracy of our results on the corpus SberQuAD that was specifically collected to solve the QA problem for Russian. Finally, we trained our model on the combination of the SberQuAD and ruOHQA corpora to see whether the results improve.

To evaluate the performance of each model we have chosen two main metrics used for this task: F1 and Exact Match (EM) - and have implemented their realisation from [17]. We calculated F1 and EM on test sets from SberQuAD and ruOHQA using the *transformers.metrics* taken from the HuggingFace framework.

We used the same hyperparameter values as recommended in the HuggingFace documentation [26]:

- learning rate: $2 * 10^{-5}$
- number of epochs: 3
- batch size: 16

Table 3. Evaluation of each pipeline’s performance on SberQuAD and ruOHQA with 5.040 and 399 samples respectively. We report the exact match (EM) and F1 metrics.

DS for training	DS for test	distilrubert-tiny-cased-conversational		distilrubert-small-cased-conversational		Distilrubert-base-cased-conversational	
		F1	EM	F1	EM	F1	EM
SberQuAD	SberQuAD	52.206	33.459	52.231	33.141	78.114	58.161
	ruOHQA	54.581	26.202	30.843	56.925	63.234	38.795
SberQuAD+ruOHQA	SberQuAD	50.557	32.324	49.914	32.196	78.160	58.503
	ruOHQA	80.558	67.229	80.118	67.470	79.557	63.373

Table 3 compares the accuracy evaluation results for different combinations of train sets and three versions of the distilrubert model. The significant boost in performance appears after subjoining entries from the ruOHQA corpus to the SberQuAD dataset. The best accuracy value of F1 metric (80.558) tested on the ruOHQA dataset was achieved by the distilrubert-tiny-cased-conversational model trained on the combined SberQuAD and ruOHQA dataset. The best EM result (67.470) was achieved by the distilrubert-small-cased-conversational on the same dataset. However, we see significant differences if we compare the rates obtained for SberQuAD and ruOHQA to the lower accuracy results made only on the SberQuAD dataset. Such a high performance on the combined dataset represents an interesting finding. We might presume that the

distilrubert-tiny-cased-conversational and distilrubert-small-cased-conversational show volatility when additional data is subjoined, consequently, models outperform the distilrubert-base-cased-conversational model in values.

In view of those considerations, we can conclude that the distilrubert-base-cased-conversational model fine-tuned on the combined SberQuAD and ruOHQA dataset can be considered as the most stable and simultaneously showing decent results on both datasets: 78.160 of F1 metric and 58.503 of EM.

Additionally, worth noticing is the fact that the models trained only on the SberQuAD dataset do not show high performance when tested on the ruOHQA set. We explain this by significant differences in the data structure of the sets, since entries in the SberQuAD corpus initially existed in written form while in the ruOHQA corpus they are compiled from oral history archives, i.e. have conversational spoken form.

6 Conclusion

This article presents the results of training the QAS models on ruOHQA and SberQuAD datasets. The content of our collected corpus initially has an oral form and is largely influenced by the emotional state and age of the respondents. Since training QAS requires structured training data, the ruOHQA corpus was annotated not only automatically, but also manually. In our paper, we described the method we followed to carry out the tagging. In addition, we presented some statistical characteristics of the resulting dataset.

As a result of our research, a demonstration dataset containing answers, questions and contexts based on interviews with Holocaust survivors was processed and published as a HuggingFace Dataset [27].

We used our corpus in combination with the SberQuAD dataset to conduct some experiments with three distilled ruBERT models. Incorporating of the ruOHQA dataset positively influences evaluation results. The best gotten F1 equals 80.558% reached by the distilrubert-tiny-cased-conversational model. However, our results showed that the distilrubert-base-cased-conversational model turns out to be more stable reaching appropriate F1 and EM scores at the same time. Moreover, it was found that results on RuOH-test in some situations are slightly better than those on SberQuAD-test before fine-tuning of tiny distied RuBERT, which will require deeper research. We are planning to try other training setups, for example, comparing the current setup with pretraining on SberQuAD and then fine tuning on ruOHQA, and also other state-of-the-art models, such as ELECTRA, T5 and LLMs.

In our future research, we plan to expand the ruOHQA corpus by processing other materials from oral history archives, e.g. the Shoah Foundation [3] containing about 7,000 video interviews with people who survived the Holocaust. These recordings in Russian have no annotated text presented. Our further work for this reason may include developing a speech recognition system.

References

1. Bouziane, A., Bouchiha, D., Doumi, N., Malki, M.: Question answering systems: survey and trends. *Procedia Comput. Sci.* **73**, 366–375 (2015). <https://doi.org/10.1016/j.procs.2015.12.005>

2. Abrams, L.: *Oral History Theory*. Routledge, London (2016). <https://doi.org/10.4324/978131564076>
3. USC Shoah Foundation. <https://sfi.usc.edu/>. Accessed 22 Sept 2023
4. YouTube playlist of Holocaust survivors testimonies by Yad Vashem foundation. <https://www.youtube.com/playlist?list=PLanQ0TFmIYBTv8sRAkSDWQLZNhbM-vIxp>
5. Picheny, M., Tüske, Z., Kingsbury, B., Audhkhasi, K., Cui, X., Saon, G.: Challenging the boundaries of speech recognition: the MALACH corpus. In: *Interspeech 2019*, pp. 326–330 (2019). <https://doi.org/10.21437/Interspeech.2019-1907>
6. Psutka, J.V., Pražák, A., Vaněk, J.: Recognition of heavily accented and emotional speech of English and Czech Holocaust survivors using various DNN Architectures. In: Karpov, A., Potapova, R. (eds.) *SPECOM 2021. LNCS*, vol. 12997, pp. 553–564. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87802-3_50
7. Chýlek, A., Švec, J., Šmídl, L.: Initial experiments on question answering from the intrinsic structure of oral history archives. In: Karpov, A., Potapova, R. (eds.) *SPECOM 2021. LNCS*, vol. 12997, pp. 124–133. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87802-3_12
8. Project MALACH – Multilingual Access to Large Spoken Archives. <https://malach.umiacs.umd.edu/>
9. Byrne, W., et al.: Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans. Speech Audio Process.* **12**(4), 420–435 (2004). <https://doi.org/10.1109/TSA.2004.828702>
10. Mihajlik, P., Fegyó, T., Németh, B., Tüske, Z., Trón, V.: Towards automatic transcription of large spoken archives in agglutinating languages – Hungarian ASR for the MALACH project. In: Matoušek, V., Mautner, P. (eds.) *TSD 2007. LNCS*, vol. 4629, pp. 342–349. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74628-7_45
11. Psutka, J., Ircing, P., Psutka, J.V., Hajič, J., Byrne, W., Mírovský, J.: Automatic transcription of Czech, Russian and Slovak spontaneous speech in the MALACH project. In: *Eurospeech 2005*, pp. 1349–1352. ISCA (2005). <https://doi.org/10.21437/Interspeech.2005-489>
12. Ramabhadran, B., Huang, J., Picheny, M.: Towards automatic transcription of large spoken archives - English ASR for the MALACH project. In: *ICASSP 2003*, p. I (2003). <https://doi.org/10.1109/ICASSP.2003.1198756>
13. Ramabhadran, B., et al.: USC-SFI MALACH interviews and transcripts English. In: LDC2012S05. Web Download. Philadelphia: Linguistic Data Consortium (2012). <https://doi.org/10.35111/7zfn-a492>
14. Psutka, J., Radová, V., Ircing, P., Matoušek, J., Müller, L.: USC-SFI MALACH interviews and transcripts Czech. In: LDC2014S04. Web Download. Linguistic Data Consortium, Philadelphia (2014). <https://doi.org/10.35111/v2nt-7j09>
15. Chýlek, A., Šmídl, L., Švec, J.: Question-answering dialog system for large audiovisual archives. In: Ekštejn, K. (ed.) *TSD 2019. LNCS*, vol. 11697, pp. 385–397. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27947-9_33
16. Eřimov, P., Chertok, A., Boytsov, L., Braslavski, P.: SberQuAD – Russian reading comprehension dataset: description and analysis. In: Arampatzis, A., et al. (eds.) *CLEF 2020. LNCS*, vol. 12260, pp. 3–15. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_1
17. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, November, pp. 2383–2392. Association for Computational Linguistics (2016). <https://doi.org/10.48550/arXiv.1606.05250>
18. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: unanswerable questions for SQuAD. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia, pp. 784–789. Association for Computational Linguistics (2018). <https://doi.org/10.18653/v1/P18-2124>

19. Pisarevskaya, D., Shavrina, T.: WikiOmnia: filtration and evaluation of the generated QA corpus on the whole Russian Wikipedia. In: Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Abu Dhabi, United Arab Emirates (Hybrid), pp. 125–135. Association for Computational Linguistics (2022). <https://doi.org/10.18653/v1/2022.gem-1.10>
20. Morphological analyzer pymorphy2. <https://pymorphy2.readthedocs.io>
21. NLTK documentation. <https://www.nltk.org>
22. Kolesnikova, A., Kuratov, Y., Konovalov, V., Burtsev, M.: Knowledge distillation of Russian language models with reduction of vocabulary. In: Proceedings of the International Conference «Dialogue 2022», Moscow, 15–18 June 2022. Computational Linguistics and Intellectual Technologies, vol. 21, pp. 295–310 (2022). <https://www.dialog-21.ru/media/5770/kolesnikovaaplusetal036.pdf>. ISBN 978-5-7281-3205-9
23. Wolf, T., Debut, L., Sanh, V., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics (2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
24. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2019). arXiv [arXiv:1910.01108](https://arxiv.org/abs/1910.01108). <https://arxiv.org/abs/1910.01108>
25. Kuratov, Y., Arkhipov, M.: Adaptation of deep bidirectional multilingual transformers for Russian language, arXiv preprint (2019). [arXiv:1905.07213](https://arxiv.org/abs/1905.07213). <https://arxiv.org/abs/1905.07213>
26. Question Answering, HuggingFace documentation. https://huggingface.co/docs/transformers/tasks/question_answering
27. RuOHQA dataset on HuggingFace. https://huggingface.co/datasets/Mihaj/ruohqa_demo