Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 508–514
Marseille, 20-25 June 2022

Russian Jeopardy! Data Set for Question-Answering Systems

Elena Mikhalkova, Alexander Khlyupin
Tyumen State University
625003, Volodarskogo, 6, Tyumen, Russia
e.v.mikhalkova@utmn.ru

Abstract

Question answering (QA) is one of the most common NLP tasks that relates to named entity recognition, fact extraction, semantic search and some other fields. In industry, it is much valued in chat-bots and corporate information systems. It is also a challenging task that attracted the attention of a very general audience at the quiz show Jeopardy! In this article we describe a Jeopardy!-like Russian QA data set collected from the official Russian quiz database Chgk *che-ge-'ka:*. The data set includes 379,284 quiz-like questions with 29,375 from the Russian analogue of Jeopardy! – "Own Game". We observe its linguistic features and the related QA-task. We conclude about perspectives of a QA challenge based on the collected data set.

Keywords: question answering, open-domain, quiz, Jeopardy!, Own game, Chgk, corpus, challenge, evaluation

1. Introduction

In natural language processing (NLP), question answering (QA) is one of the most common tasks that encompasses a number of question types, including "questions about everything", the so-called open-domain QA (Chen and Yih, 2020). Open-domain questions cover a wide range of topics and do not necessarily come in form of an actual question (e.g. "Who is the living Queen of England?") which draws the task of answering them very close to information retrieval. The query can be just a line of keywords: *living Queen England*, but pragmatically it is still a question. From this broad perspective, QA is developed in production of search engines, corporate information systems and conversational technologies like chat-bots.

In February 2011, Watson, an IBM's information system (Ferrucci et al., 2010) installed in a small computer, won against two very prominent human players in a TV quiz-show called Jeopardy! [1] The algorithm was trained on TREC corpus (Voorhees, 1999) and 500 questions manually collected from the TV-show. In TREC, questions are formulated quite typically, e.g. "How many calories are there in a Big Mac?", although they cover a variety of topics. In contrast to it, the Jeopardy! challenge presents questions as clues narrowed by a certain domain like in the following example from (Ferrucci et al., 2010):

Category: Oooh. . . .Chess
Clue: Invented in the 1500s to speed up the game, this maneuver involves two pieces of the same color.
Answer: Castling

The existing open-source Russian QA data sets are more like trivia questions and answers resembling TREC: RuBQ (Korablinov and Braslavski, 2020) consists of 1,500 Russian questions loaded from various "quiz collections on the Web" with answers linked to Wikidata entities; RuBQ 2.0 has "2,910 questions along with the answers and SPARQL queries" (Rybin et al., 2021); SberQuAD (Efimov et al., 2020) contains "50,364 paragraph–question–answer triples" that are now publicly available; the questions were written by crowd-annotators.

In this article, we observe a data set of Russian Jeopardy! questions and answers and outline a related QA challenge. The database of questions and answers called Chgk *che-ge-'ka:* is freely available at `https://db.chgk.info/`. Our current contribution includes the following:

1. We describe the Russian Chgk QA database containing nearly 400K questions and answers that test players' logic and erudition.
2. We describe its sub-corpus of Jeopardy!-like questions and outline its characteristic features.
3. We formulate a QA-challenge based on the Russian Jeopardy! data set.

2. Russian ChGK Database

There exists a variety of Russian intellectual games (quizzes) some of which have formed very devoted communities in and even outside Russia. "What? Where? When?" (*Chto? Gde? Kogda?*, hence the abbreviation Ch-G-K) is one of the most popular Russian TV quiz shows, dating back to 1975 [2]. As the TV game show allows but a few players (a team of six) per one episode, in the 1990s the game format spread among common people who wrote questions and played them at local ChGK tournaments. The movement grew into the so-called "Sports Chgk". The community site that collects information about the movement [3] contains ratings of 228,438 players from 54,995 teams (as of 7 May 2022). ChGK tournaments

[1] `https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html`

[2] `https://en.wikipedia.org/wiki/What%3F_Where%3F_When%3F`

[3] `https://rating.chgk.info/`

are organized in Montreal, Richmond Hill, Vilnius, Odessa, Kharkiv, Cologne, Boston, Nahariya, Eilat, Parnu, Astana, Vladivostok and many other cities all over the world. The movement has an official open access collection of about 400K questions in Russian. In view of its size, metadata, effort of the community that supports it, this database can be considered cultural heritage of the Russian language. The earliest tournament in the database is the 1990's "I Championship MAK in "What? Where? When?" 1990-01-01". [4] The copyright allows to use it for non-commercial purposes with some of the packs (collections of questions played during one tournament) distributed under different Creative Commons licenses [5]. Packs are written, tested, approved and then played at different events (offline and online) under different commercial and non-commercial terms. After a row of tournaments, packs are uploaded to the database. Amateur and semi-professional packs usually do not go to the official database. The moment the packs are uploaded they are under the database license.

Due to the number of people involved in the movement and not only for sports, but commercial interest, we can say that for some people sports ChGK is a profession. Our experience of communicating with the community shows that writing questions for ChGK is demanding and depends on authors' reputation. There are well-known authors who earn money by writing and testing questions, and entrepreneurs who organize commercial tournaments. Hence, we can say that the ChGK database contains *professionally written* questions, maybe, not from the beginning of the 1990s, but from the times when it became business for many authors and organizers.

As the website allows only specific search in the database, we parsed the XML-tree of tournaments at `https://db.chgk.info/tour` with the Python library BeautifulSoup [6] and gathered all the QA information from HTML-pages. The general metadata include: Question, Answer, Author, Sources (Web-links that authors used to write a question), Comments (by authors and organizers), Pass Criteria (in case players' answers are not very precise), Notices (comments by players), Images (Web-links to pictures if they are needed in a question), Rating (hardness of the question calculated from how many teams managed to answer it), Number of question (in order in each pack), Tournament type. The metadata for Jeopardy! questions also include Topic (a common topic for a set of 5 questions, traditionally called "a category") and Topic Number (in the order of sets of questions from one tournament). The data were collected in form of .csv tables locally and uploaded to our own SQL-database (see a part of its scheme referring to the corpus in Fig. 1).

[4] `https://db.chgk.info/tour/mak1`

[5] `https://db.chgk.info/copyright`

[6] `https://www.crummy.com/software/BeautifulSoup/bs4/doc/#`

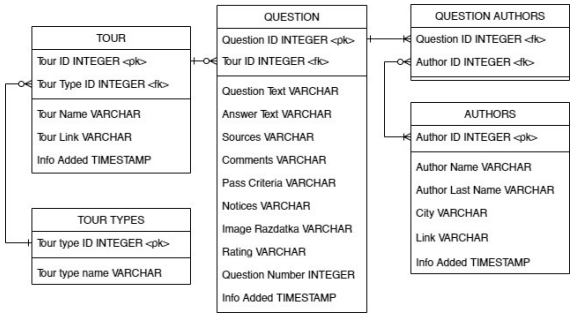

Figure 1: A part of our downloaded ChGK database illustrating metadata about questions.

As mentioned, the sports ChGK includes questions of different types depending on the type of tournament that they are played at. In total, there are nine types: author's, championships at different countries and regional tournaments (the format can be purely of its authors' design, although it usually complies with the general style), "synchrons" (typical ChGK questions played at tournaments simultaneously by many teams), Internet and television quizzes, questions for training, topical questions, questions in a poetic form (verses), questions for erudition (i.e. based purely on knowledge) and of the Jeopardy! type. The most of the database resembles the following example.

Question 7, the tournament "ChGK is... - 2017": [7]
The legend has it that once Paul Bunyan fired his gun at a deer and ran to get his prey. But he ran so fast that he DID THIS and felt an itch in his back. What did he do?

Answer: He outran the bullet.

This question was played at tournaments for teams of two players, usually held on the 14 of February. Hence, its title, that resembles the name of the chewing gum "Love is..". It is supposed to be an easy question so that a small team can solve it; experienced players would consider it straightforward, giving out a lot of details to find the correct answer. Questions in tournaments for teams of six are a lot more obscure: they give too many or too few details, contain misleading metaphors.

The question is written according to a very common formula: "DID THIS". Often the question-like part ("What did he do?" in the example above) is omitted. At tournaments, the host reading it would put an additional stress on the phrase "DID THIS", so that the players understand that this is a question in form of a statement. In the database, such phrases are italicized or capitalized like in our example.

Although the question mentions the detail – the American folk hero Paul Bunyan and a deer – the answer does not require this information. It can be derived only from the situation with the bullet, running and an itch

[7] Authors V. Ostrovskiy, A. Boyko, M. Podryadchikova. Translation into English is ours. `https://db.chgk.info/tour/eila08a1.2`

in the back. This is a way of misleading players that grab at several hints and need to choose the correct semantic, logical and factual track that narrows the choice of answer. It is important that answers derived from wrong tracks should be incoherent or contradict some facts that are omitted in the question, so that the correct answer cannot be criticized. In the classification offered by (Dimitrakis et al., 2020) [8], such questions are called *procedural*.

The ChGK database also contains questions of Jeopardy! type. In Russian, Jeopardy! is called "Own Game" *Svoya Igra.* [9] Nearly all these questions are in form of statements and the object of interest, about which the question is asked, is often capitalized. Let us study the following example.

Topic 7: Parrots. Question No. 3. [10] The last name of THIS famous DETECTIVE is translated as "a parrot".

Answer: Hercule Poirot.

Note that this question is shorter and more fact-oriented than the ChGK question before it. It is meant for single players competing against each other, although there are variants of "Own Game" for teams of two, three and four. The question above requires to compare two rows of data: words denoting "parrot" in different languages and last names of famous detectives. In the classification by (Dimitrakis et al., 2020), this type of questions is called factoid, meaning that it resembles a fact, but it is missing some information. Some ChGK questions are also very close to this type, especially in tournaments called *lite*, i.e. easier tournaments for new-comers and younger players.

Like in Jeopardy!, questions in "Own Game" are organized according to topics (categories) in packs of five, from the easiest to the hardest. The number of the question in the example above is 3 which denotes that it is of medium hardness. At the sports "Own Game", i.e. not the television version, the player who answers it faster than others will get 30 points (the easiest question weighs 10 points and the hardest – 50).

Often the topic is a direct hint to the answer, so adding topics to a QA system's input is supposed to be useful. Consider the following question:

Topic: OST. **Question:** Van Gogh, Gauguin, and Toulouse-Lautrec belonged to THIS movement. **Answer:** Post-impressionism. [11]

"OST" means that this combination of letters should be in the answer, which leads to just "impressionism" not being the answer. In the following case, it is impossible to derive the question without knowing the topic:

Topic: Authors of questions. **Question:** Are you jealous? **Answer:** Paul Gauguin. [12]

The topic "Authors of questions" presupposes that players should remember famously known questions like the one which is the title of Gauguin's picture.

Although usually fact-oriented, the Russian Jeopardy! questions can be of the logical type as well:

Topic: Don Aminado. **Question:** Of the two who are going to bet, the both risk: one – to lose, the other – ... **Answer:** To never be paid. [13]

The question above is based on a quotation by a famous Russian writer and entails knowledge of a real-life situation, although questions of this type are not very common in the Russian Jeopardy! database. The game does not only check who knows more facts and can recall them faster than others. It also checks how accurate players are when they evaluate possibility that the answer they have just come up with is correct. Rare logical questions based on common sense and typical situations, evidently, aim at the latter skill.

It is also important that some questions in the Chgk database depend on additional media:

1. images (but not videos; however, this is not the case with the TV game show) which can not only be photos, screenshots, etc., but visual aids like schemes, symbols, texts printed on handouts (called "razdatka" and named so in our scheme 1);

2. the host's intonation. Comments to questions often contain remarks on how to pronounce some parts, for example, without giving out the answer. Intonation can also be marked with capital letters.

3. Russian Jeopardy! Data Set

As mentioned, we downloaded data from the official sports ChGK resource to be able to parse them and store in different formats. The Russian Jeopardy! (*Own Game*) data set seems to us to be the most valuable for NLP as:

1. its questions are shorter than in other quizzes and more fact-oriented;

2. it is quality-guaranteed, as it was created by professional authors;

3. it is suitable for open-domain QA;

4. it has additional information like links to Web-sources and question ranking that points at its "hardness";

5. above all, it is not too trivial in the field of QA data sets and hence it can foster new tasks and approaches in QA itself.

[8] Which they attribute to (Mishra and Jain, 2016), but the classification by Mishra and Jain (2016) is only a part of their list.

[9] https://www.imdb.com/title/tt1381017/

[10] Author Oleg Sarayev. Translation into English is ours. https://db.chgk.info/tour/eu05stsv

[11] Author Yuri Grishov. Translation into English is ours. https://db.chgk.info/tour/grishov

[12] Author Yuri Grishov. Translation into English is ours. https://db.chgk.info/tour/grishov

[13] Author Yuri Grishov. Translation into English is ours. https://db.chgk.info/tour/grishov

The last point is more vividly discussed by (Boyd-Graber and Börschinger, 2020).

Currently, the data set (Mikhalkova, Elena and Alexander Khlyupin, 2021) contains 29,375 questions from the ChGK database. The questions were selected based on the following criteria:

1. a question is in form of a text;
2. a question does not have an image supporting it;
3. a question does not mention that any images should be distributed or shown on a screen while solving it.

I.e. these are fully verbalized questions. The data set including some "flattened" metadata from the database scheme 1 is stored in a .csv file at `https://github.com/evrog/Russian-QA-Jeopardy`. The delimiter is tabulation. The data include: Question ID, Question, Answer, Topic, Authors' Full Names, Name of tournament, Link to Tournament. The rest of the information supplying questions has not been included in the data set as it is not given to players during the game, but it is available via links to tournaments.

Table 1 gives a summary of the whole ChGK data set, as downloaded on 6 July 2021 and updated on 18 November 2021. In the table, "Synchron" is a typical ChGK tournament played immediately by several teams of six players maximum; "Lite" is its mentioned version with easier questions. Its questions are shorter, but unlike Jeopardy! they are more logic-oriented. The table demonstrates that Jeopardy! questions are twice shorter in length than typical ChGK questions. And even lite questions are not near them in length.

Table 2 describes distribution of words of different parts of speech across the Jeopardy! data set, classified with the help of NLP-software spaCy (Honnibal et al., 2019). It is of no surprise that nouns are the most frequent category, but, among them, proper nouns stand out. Proper nouns are much less frequent in topics ($\tilde{1}6\%$) and questions ($\tilde{1}7\%$). However, they are in $\tilde{3}4\%$ of answers. NER-classifier by spaCy defines that persons, organizations and locations are approximately equally distributed in questions. However, in answers persons comprise as many as 70% of the classified entities.

It is also natural that verbs are about three times more frequent in questions, than in answers and topics. The most frequent verb is "to name" (3,607 tokens), probably, due to a typical formula of a question "Name somebody or something that.." which is a variant of "THIS somebody or something..". As for other actions expressed by verbs, beside "being" or "becoming" and their variants, they are "to gain" (783 tokens), "to wear" (649), "to write" (546), "to have" (425), "to be located, situated" (411), "to say" (360), "to call" (360), "to tell" (348), "to belong" (334), "to mean" (333), "to do" (295), "to play" (284), "to write" (275), "to be considered" (266), "to happen" (263), "to create" (257), "to paint, describe" (242) etc. These verbs hint at a more general topic, for example, art, poetry, music, sports, awards, famous quotes. Although, due to polysemy, the verb "to gain" is used in quite a variety of topics.

4. Task Discussion

In this paragraph, we describe a challenge based on the Russian Jeopardy! data set. The challenge will be held in two formats: online and offline. The online format will be supervised by the team of the project Russian SuperGLUE (Shavrina et al., 2020). The task will appear in the project's online system around June 2022. [14] The offline format is organized as a series of Jeopardy! games at the Tyumen State University where QA systems will compete against actual players, and the first game is scheduled in July 2022. We further describe the offline format, as our team is responsible for it.

4.1. Jeopardy! Game Format

Following the Jeopardy! challenge of February 2011, mentioned earlier, we propose that at our event two experienced players compete against one system. At the first challenge, we will test three different systems in two rounds of five topics (categories), the first one containing easier questions and the second – harder. Players will change after two rounds, too, i.e. each system will be competing with two new players. The classic Jeopardy! also consists of two sets of categories, probably, because more rounds would wear players out.

As we do not test QA systems' acoustic technologies, during the game each system only needs an interface for texting which will be supervised by an operator. This interface, be it a command line or graphic user interface, will be broadcast on a screen behind players. When a new question is opened and the host starts reading it, a game manager will send the question in textual form to the system's operator. When the system returns the answer, its operator will press the signalling button. If he or she does it before other two players, the operator will read the answer, and the host will evaluate it as correct or incorrect. The operator is not allowed to change the system's answer, but he or she can abstain from pressing the button.

The rest of the game rules coincide with the classic version. [15] Hence, the task for QA systems is to automatically answer as many questions as possible, as correctly as possible, and as fast as possible. It is advisable that systems weigh their confidence before they return the answer, so as not to lose points on nonsensical answers. However, at our first game operators have the right to not press the button and prevent such cases.

[14] `https://russiansuperglue.com/tasks/`

[15] See, for example, its layout in Wikipedia `https://en.wikipedia.org/wiki/Jeopardy!#Gameplay`.

Type	Questions	Tours	Average Q length in tokens	Average Q length in symbols
Jeopardy!	29,375	452	14.28	98.37
ChGK Synchron	48,065	1,821	32	234
ChGK Lite	1,936	54	27.5	201
All	379,284	4,816	34	244.9

Table 1: Details about the sports ChGK database, as of 6 July 2021 and partially updated 18 November 2021. Tours – tournaments; Q – question.

Part-of-speech	No. of words	%
Questions		
All nouns:	160,844	62.12
Regular nouns	116,945	45.17 (72.71)*
Proper nouns	43,899	16.95 (27.29)
Persons**	19,551	7.55 (46.61)***
Organizations	12,123	4.68 (28.9)
Locations	10,268	3.97 (24.48)
Verbs	49,671	19.18
Adjectives	48,407	18.70
Total	**258,922**	-
Answers		
All nouns:	55,642	79.43
Regular nouns	31,914	45.56 (57.36)
Proper nouns	23,728	33.87 (42.64)
Persons	13,579	19.39 (70.26)
Organizations	2,970	4.24 (15.37)
Locations	2,777	3.96 (14.37)
Verbs	4,704	6.72
Adjectives	9,702	13.85
Total	**70,048**	-
Topics		
All nouns:	36,767	77.25
Regular nouns	29,299	61.56 (79.69)
Proper nouns	7,468	15.69 (20.31)
Verbs	3,411	7.17
Adjectives	7,415	15.58
Total	**47,593**	-

Table 2: Distribution of parts-of-speech in questions, answers and topics of Jeopardy! data set. *For regular and proper nouns, numbers in round brackets denote percentage among all nouns. **For proper nouns, persons, organizations and locations were derived with spaCy; other entities have not been classified. ***Percentage among all defined entities.

4.2. Test Set

Currently, there is one open access data set for the project – the one we described in the previous paragraph. It has not been split into training and developer sets, as it is common to use Web-connection in QA systems and ChGK questions are easily found on the Web with the help of search-engines. For the online system at Russian SuperGLUE, several ChGK authors created a closed test set of 512 questions. This test set will be placed in the system to evaluate online submissions.

As for the set at the offline game, it will also be a pack of yet unpublished questions written by authors of sports ChGK. These questions will be only textual, with no visual or audio support, and with as little metaphoricity and wordplay as possible. For each set of two rounds the questions will be organized in six topics (categories) and distributed into easier and harder. After the questions will be played at our first game, they will be added to our data set as a developer set.

4.3. Baseline

As the baseline for our project, we suggest the open-domain question-answering model for Russian based on Wikipedia, developed by DeepPavlov project: `odqa.ru_odqa_infer_wiki`. The starter code and its description can be found here `https://docs.deeppavlov.ai/en/master/features/skills/odqa.html`. To help developers install the baseline in Google Collaboratory, we have added a Jupyter Notebook in our mentioned repository `https://github.com/evrog/Russian-QA-Jeopardy`. We have tested our baseline on the first 800 questions in our data set, and manually checked correctness of answers. The result is 16 correct answers, i.e. approximately 2 per 100. We tried manual rephrasing questions into a question-like form, and it helped to get some answers right, but just in a few cases. Also, we tried adding a topic to a question and it lowered the performance from 16 to 12 correct answers. Hence, in the current version DeepPavlov cannot compare to actual players and needs training and facilitation before it is ready for competition.

4.4. Evaluation

At the offline event the host evaluates correctness of the answer and game managers keep the score. However, for system developers we suggest that to prepare their systems they can consider the following. The evaluation stage consists of two steps: the first metric compares the answer to the correct one, the second metric calculates the system's performance. The both metrics vary across existing QA projects.

The first metric varies as correctness of the answer can be understood differently. In case of the answer to the

previously discussed question, the following variants are equally possible: “Hercule Poirot”, “It is Hercule Poirot.”, “Poirot” and other versions meaningfully referring to this character and no other. As mentioned by Chen et al. (2019), many metrics used in QA evaluation are imported from machine translation. They also note that the METEOR metric (Banerjee and Lavie, 2005) is, by the result of their study, the closest to human judgments. (Niwattanakul et al., 2013; Thada and Jaglan, 2013; Rahutomo et al., 2012) consider Jaccard, Dice, Cosine Similarity and similarity distance, e.g. Damerau–Levenshtein distance (Levenshtein and others, 1966). The SberQuAD challenge of 2017 evaluated exact matches with the gold standard and overlaps of tokens averaged over all questions (Efimov et al., 2020).

We checked several metrics on the obtained answers from our baseline. The metrics were Cosine Similarity by spaCy (Honnibal et al., 2019), Jaccard distance, Damerau–Levenshtein edit distance, METEOR by NLTK (Bird et al., 2009). After we calculated the metrics for 800 answers, we sorted the range and found the number of correct answers from its top. The lowest was the result of Cosine Similarity: the first answer was incorrect, and then only one correct answer before the next error which is most likely due to absence of many words, especially proper nouns and word groups, in the model. The results of the rest of the metrics are: Damerau–Levenshtein – 5 correct answers, Jaccard – 9, METEOR – 14, which corresponds to conclusions by (Banerjee and Lavie, 2005). The mean between the rounded METEOR coefficient for the last correct answer in the ranged set and for the incorrect answer next to it is $(0.238095 + 0.217391)/2 = 0.227743$ which we, currently, propose as the minimum to automatically evaluate answers as correct.

As for the second metric, Calijorne Soares and Parreiras (2020) enlist the usual NLP measures such as Precision, F1 as well as less known Mean Average Precision (MAP), used in Information Retrieval. However, it is also possible to modify the scoring system of Jeopardy! to evaluate developed systems. In Jeopardy!, giving a correct answer adds points, giving an incorrect answer subtracts points, and giving no answer doesn't affect the score. Also, as mentioned, questions in each topic are organized from the easiest to the hardest. Hence, the score for the question depends on its rank in the topic.

For a given answer a, its rank r in a topic t is the order of the question, to which the answer is given: r_t.

$$f(a) = \begin{cases} a = r_t & \text{if answer is correct} \\ a = 0 & \text{if no answer is given} \\ a = -r_t & \text{if answer is incorrect} \end{cases} \quad (1)$$

The result of the system is the sum of its scores for a set of n questions: $\sum_{i=1}^{n} a_i$. The evaluation code can be found in our GitHub repository. To compare to earlier versions and other systems, the score should be evaluated on test sets of the same size.

4.5. System Description

After each of our offline challenges, we will publish their results and description of participating systems. We consider it vital that system description directly states whether the system searches for the answer in the Internet, as the task is obviously harder to solve without Web-connection – how actual players do during the game. Also, system description should clearly state which already existing software and Web-services (including search engines) the system uses beside original tools.

5. Conclusion

The paper describes the database of the Russian professional quiz-writers, ChGK, that contains 379,284 questions, answers and other metadata, like links to sources of questions. We organized its sub-set into the first data set of 29,375 Russian Jeopardy! (“Own Game”) questions and answers. We touch upon types of ChGK tournaments and analyze several examples of Jeopardy! questions. We also describe several statistical features of the data set. Finally, we outline the Russian Jeopardy! QA challenge that will be held in summer 2022 at the Tyumen State University: we touch upon our motivation, the game format, test set, baseline and evaluation metrics.

As mentioned earlier, the Jeopardy! data set written for tournaments by professional authors is not a trivial set of questions and poses unusual tasks for the QA field. We are hopeful that our proposed challenge will bring valuable results in Russian QA, and we also plan to continue it with the logical, and not only fact-oriented, type of questions.

6. Bibliographical References

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. ”O'Reilly Media, Inc.”.

Boyd-Graber, J. and Börschinger, B. (2020). What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online, July. Association for Computational Linguistics.

Calijorne Soares, M. A. and Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646.

Chen, D. and Yih, W.-t. (2020). Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37.

Chen, A., Stanovsky, G., Singh, S., and Gardner, M. (2019). Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124.

Dimitrakis, E., Sgontzos, K., and Tzitzikas, Y. (2020). A survey on question answering systems over linked data and documents. *Journal of Intelligent Information Systems*, 55(2):233–259.

Efimov, P., Chertok, A., Boytsov, L., and Braslavski, P. (2020). Sberquad–russian reading comprehension dataset: Description and analysis. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–15. Springer.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., et al. (2010). Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79.

Honnibal, M., Montani, I., Honnibal, M., Peters, H., Landeghem, S. V., Samsonov, M., Geovedi, J., Regan, J., Orosz, G., Kristiansen, S. L., McCann, P. O., Altinok, D., Roman, Howard, G., Bozek, S., Bot, E., Amery, M., Phatthiyaphaibun, W., Vogelsang, L. U., Böing, B., Tippa, P. K., jeannefukumaru, GregDubbin, Mazaev, V., Balakrishnan, R., Møllerhøj, J. D., wbwseeker, Burton, M., thomasO, and Patel, A. (2019). explosion/spaCy: v2.1.7: Improved evaluation, better language factories and bug fixes, August.

Korablinov, V. and Braslavski, P. (2020). Rubq: a russian dataset for question answering over wikidata. In *International Semantic Web Conference*, pages 97–110. Springer.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10:8, pages 707–710. Soviet Union.

Mishra, A. and Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1:6, pages 380–384.

Rahutomo, F., Kitasuka, T., and Aritsugi, M. (2012). Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*, volume 4:1, page 1.

Rybin, I., Korablinov, V., Efimov, P., and Braslavski, P. (2021). Rubq 2.0: An innovated russian question answering dataset. In *European Semantic Web Conference*, pages 532–547. Springer.

Shavrina, T., Fenogenova, A., Anton, E., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., and Evlampiev, A. (2020). RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online, November. Association for Computational Linguistics.

Thada, V. and Jaglan, V. (2013). Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, 2(4):202–205.

Voorhees, E. (1999). The trec-8 question answering track report. In *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8)*.

7. Language Resource References

Mikhalkova, Elena and Alexander Khlyupin. (2021). *Russian Jeopardy! Data Set*. Distributed under License https://github.com/evrog/Russian-QA-Jeopardy/blob/main/LICENSE, ISLRN 842-398-662-380-6.