YoGa QA - Project Report

Tsimafei Prakapenka, Seminar BERT & Friends, September 2022.

Abstract

Open-Domain Question Answering is a type of language tasks, asking a model to produce answers to factoid questions in natural language. Nowadays, research in this area is mostly focused on English language. However, there are plenty of non-English data available. For instance, there is a Russian analogue of Jeopardy called ’Svoya igra’ (Own Game).

In this paper, I independently crafted a dataset of Russian Jeopardy-like questions, initially collected by Mikhalkova (2021). Additionally, I fine-tuned mT5 model (Xue et. al, 2021) to predict an answer given question text. I demonstrate that general purpose neural language model is not enough for answering Jeopardy-like questions in Russian. In the future work, I propose to use a combination with structural knowledge (knowledge graphs) or extend neural model with retrieval component (Wikipedia passages). All data, code and experiment results are publicly available - https://github.com/tsimafeip/yoga-qa.

1. Introduction

Open-domain Question Answering (ODQA) is a type of language tasks, asking a model to produce answers to factoid questions in natural language. The true answer is objective, so it is simple to evaluate model performance (definition by Weng, 2020).

State-of-the-art architectures for this task are neural-based models with an external knowledge index, assessed by retriever component. The most recent ATLAS model (Izacard, 2022) and its predecessor RAG (Lewis et al., 2020) are both based on two sub-models: the retriever and the language model. When generating an answer, the model starts by retrieving the top-k relevant documents from a large corpus of text with the retriever. Then, these documents are fed to the language model, along with the query, which, in turn, generates the output. Both the retriever and the language model are based on pre-trained transformer networks.

Speaking about evaluation of question answering models, there are several benchmarks available. In the paper body, I will describe three of them: Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017) and SearchQA (Dunn et al., 2017). Importantly, all of these datasets are in English, so questions

answering datasets in other languages are severely underrepresented in the current QA research. Hence, moving from English-centric QA to a multilingual one is an interesting and open research question.

Contributions of this paper are manifold:

- First, this paper describes a new edition Jeopardy-like questions in Russian, initially presented in Mikhalkova (2021). I independently built a dataset from the same source, but with an improved parsing algorithm, which allowed to extract new meta information. For instance, my dataset has additional correct answers and hard negatives, specified by the author of question.
- Second, I fine-tuned mT5 model (Xue et. al, 2021) to predict an answer having only a question text, or question text and topic name. This paper contains an analysis of model performance, showing that sole neural language model is not enough for this task. I also propose a promising direction for future research: adding structured data (knowledge graphs, Wikipedia passages) to make answer prediction more accurate, grounded and interpretable.

Plan of the paper. After reviewing major QA datasets and introducing Jeopardy and Own Game concepts in section 2, I proceed to corpus creation details in Section 3. I explain the training process for neural model and discuss obtained results in Section 4. Finally, I sum up contributions of this paper and propose future directions for this specific task in Section 5.

2. Overview of QA Datasets

Modern questions answering models are competing on various benchmarks. I will describe three of them: Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017) and SearchQA (Dunn et al., 2017).

NaturalQuestions (NQ) is a dataset, consisting of real anonymised, aggregated queries issued to the Google search engine. An annotator was presented with a question along with a Wikipedia page from the top 5 Google search results, and annotates a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no long/short answer is present.

TriviaQA contains over 650K question-answer-evidence triples, that are derived by combining 95K Trivia enthusiast authored question-answer pairs with on average six supporting evidence documents per question. Evidence documents were collected *retrospectively* from Wikipedia and the Web.

SearchQA is a dataset of questions from Jeopardy - famous American quiz show. Authors start from an existing question-answer pair, crawled from J! Archive, and

augment it with text snippets retrieved by Google. Following this approach, they built a dataset, which consists of more than 140k question-answer pairs with each pair having 49.6 snippets on average.

Jeopardy is an unusual format that consists of trying to guess an entity from a fact about that entity. For example, “The World Cup” is the answer to the question “In 1986 Mexico scored as the first country to host this international sports competition twice.” Jeopardy questions are precise, factual statements, so it is easy to evaluate correctness of the generated answer by exact match.

‘Svoya Igra’ (Own Game) is a Russian analogue of Jeopardy. There is a TV version of this game, running from 1994. Additionally, there is a big community of enthusiasts in Russian-speaking countries, who write their owns questions and run championships. There is an official web database of the questions, written by professional authors - https://db.chgk.info/. There is a dataset of Own Game questions, introduced by Mikhalkova (2021). I will bring more details about the existing dataset, and Own Game specifics in the next section.

3. Dataset Collection

3.1 Own Game

Format of the data is the following: questions are grouped to the topics consisting of 5 questions somewhat related to the topic with an increasing complexity (from 1 to 5). Question context is usually around 1-3 sentences, sometimes requiring multiple assumptions to arrive to the answer. Answers are short factoid statements (≤5 words).

Topic names are usually something, that connect all questions in the topic. For instance, if topic is named ‘Dances’, then all questions will be somewhat related to different types of dances, for instance, tango or foxtrot. However, question with an answer ‘Edgar Degas’, who is a French painter, well-known for pictures of dancing woman, is a perfectly fine answer for topic called ‘Dances’. Moreover, you can have the same name of the topic, like ‘History’, ‘Films’ or ‘Cities’, but actual questions will be totally different depending on author’s knowledge and interests. There is also a special kind of topics, called ‘matrix’. It means that all answers will contain a specific substring, like ‘mor’ or ‘uri’. To sum up, topic is something common for all 5 questions, but this common trait could be literal (for example, ’Cities’), formal (’matrix’ topics) or even metaphoric (quotes from poems or songs).

For example, there is topic called ‘Switzerland’:

Question 1: ON THIS DATE in 2014, the CERN website briefly switched to Comic Sans.

Answer: 1 april (fools' date)

…… (I omit questions 2-4 for brevity sake)

Question 5: [Presenter note: do not mention quotes.] In 1963, Niklaus Wirth defended his dissertation on "Euler", but he became famous seven years later thanks to "THIS SCIENTIST".

Answer: Pascal (Wirth developed the Pascal programming language).

Both questions are requiring logic reasoning skills, but for the question about CERN you need only now a date, tied with joking. On the other hand, for the question about Niklaus Wirth, you need to guess that Pascal is not a mathematician in this case, but a programming language. Additionally, it could be useful to know that Niklaus Wirth was a computer scientist, a prominent designer of programming languages. Thus, Own Game questions are often require not only fact knowledge, but also reasoning skills.

Questions are also designed for different purposes, there are separate challenges for children, students and adults, so different datasets have different complexity in human perception. Different authors have different styles as well. For instance, one author can estimate fact complexity as 5, but another will rate it only as 1. In my opinion, checking how human perceived complexity correlates with machine results is an interesting point for the future research.

3.2 Dataset Collection

The data base of questions and answers for Own Game is freely available at https://db.chgk.info/. Data is posted as html pages, so it can be extracted by scraping and parsing. Alternatively, there are a lot of private unpublished tournaments, which are distributed in form of text files (pdf, docx, txt). However, private data is protected by author rights, so I will focus on the public database.

I used the following algorithm to extract Own Game data:

1. Collect all available subpages with valid tournaments, starting from seed page.
2. Having tournament page url, I select only tournaments with available txt version of the page. Txt files are easy to store, and format is unified for different tournaments, which is not the case for raw html pages.
3. Download txt files from web to local storage to accelerate further processing.
4. Extract data from txt file to the structured Pandas DataFrame. Parsing process is rather complicated, and has a lot of manual fixes to resolve complex cases. You can check source code, if needed.

Resulting data frame has the following required columns: topic_name, question_value, question_text, answer. There are also some optional fields: extra_positives, hard_negatives, comment, source, author, tournament, date, source_url. There is a sample dataset entry:

topic_name	question_value	question_text	answer
Океаны	*1*	*Океан, в отличие от своих братьев, не участвовал В НЕЙ, благодаря чему сохранил свое положение, а не был низвергнут в Тартар.*	*титаномахия*
Oceans	*1*	*The ocean, unlike its brothers, did not participate in IT, thanks to which it retained its position, and was not thrown into Tartarus.*	*Titanomachy*

extra_positives	hard_negatives	comment	source
битва титанов и богов	*гигантомахия*	-	*https://ru.wikipedia.org/wiki/Океан_(мифология)*
Battle between titans and gods	*Gigantomachy*	-	*English analogue: https://en.wikipedia.org/wiki/Oceanus*

author	tournament	date	source_url
Иделия Айзятулова, Андрей Мартыненко, Александр Рождествин	*XII Кубок Европы по интеллектуальным играм среди студентов (Витебск). Своя игра*	*2016-10-28*	https://db.chgk.info/txt/eu16stsv.txt
Ideliya Aizyatulova, Andrey Martynenko, Alexander Rozhdestvin	*XII European Student Intellectual Games Cup (Vitebsk). Own game*	*2016-10-28*	https://db.chgk.info/txt/eu16stsv.txt

Comparing this specific entry with the same question from Mikhalkova (2021), I see the major difference only in two aspects. First, I do extra parsing to extract extra positives and hard negatives, while in the first edition of this dataset all this meta information is stored as the ‘answer’. Second, I add the link to source wikipedia page, specified by the author. Mikhalkova (2021) mentions, that they deliberately omit source links, since

there is no access to question source at inference. I argue that including all available information from data source is useful, since it can be used for the future research. For instance, gold passages, used by authors to derive a question, are especially valuable for training a neural retrieval component.

My dataset is a smaller than one, introduced by Mikhalkova (2021): 25k vs 29k question-answer pairs. I reckon that this difference is produced by specifics of scraping process, since I do not process tournaments without txt version of the web page. To my mind, these two datasets can be merged, leading to the bigger dataset with broader meta information included.

3.3 Dataset Analysis

Own Game dataset covers questions, written in a big data range: from 1997 till 2017. It implies different answers to the same factoid questions. For example, question 'Who is the American President now?' has different correct answer, depending on the date: 'Bill Clinton' in 1997 and 'Donald Trump' in 2017. So, knowing year of the question creation could be useful during training and at inference.

Additionally, style of the questions is changed. There are two questions from the same tournament, European Student Intellectual Games Cup, but from different years: 2005 and 2016.

1. '*That was the name of the city burnt by Olga with the help of birds.' - 'Iskorosten'* (2005, source)
2. '*The composition of the cocktail WITH THIS NAME includes American whiskey and Italian amaretto. Try sometime, do a service for me.' - 'Godfather'* (2016, source)

Question, written in 2005, requires knowledge of the specific historic fact, and the reasoning here is straightforward. On the other hand, question from 2016 does not require knowledge of the particular cocktail. Instead, it implies logic reasoning. Main hero of the 'Godfather' film and book, Vito Corleone, was an Italian American and master of various 'services'. You may draw an answer from famous quote from the 'Godfather' film: *'Some day, and that day may never come, I will call upon you to do a service for me.'.* Thus, such a great variance in question style and covered topics (history, films, music) highlights complexity of this specific dataset.

4. Model Training and Results Analysis

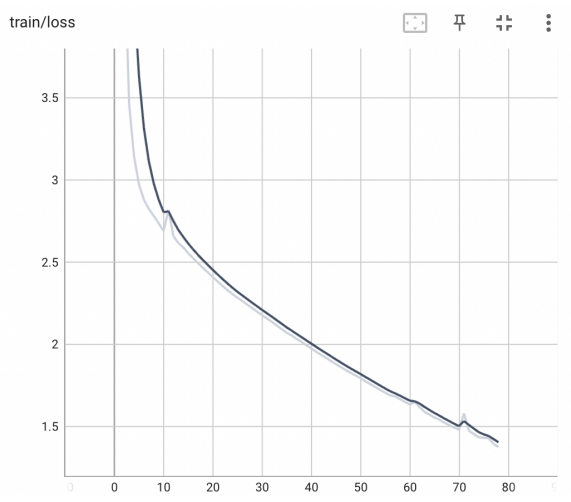
4.1 Model Training

I used mT5-base as an underlying pre-trained model. Model was fine-tuned on a single GPU Tesla V100-PCIE-32GB for 2 days. All training parameters can be found in the

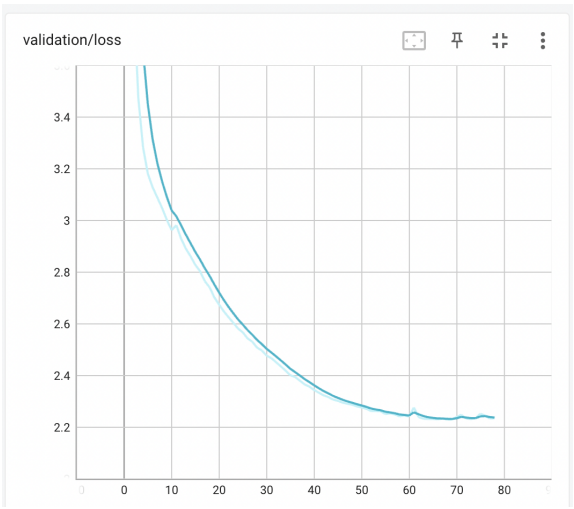
config. I have not split tournaments (packs of questions) to avoid overfitting, so every single tournament goes either to train, dev or test splits. I used AllenNLP as a framework for training.

Model used the dataset, consisting only of question-answer pairs. I also tried appending topic to the question text by simply concatenating them ('TOPIC. QUESTION_TEXT.'). This model is still training, but I expect only slight improvements in quality. To highlight, all the future analysis is for simple model fine-tuned on question-answer pairs.

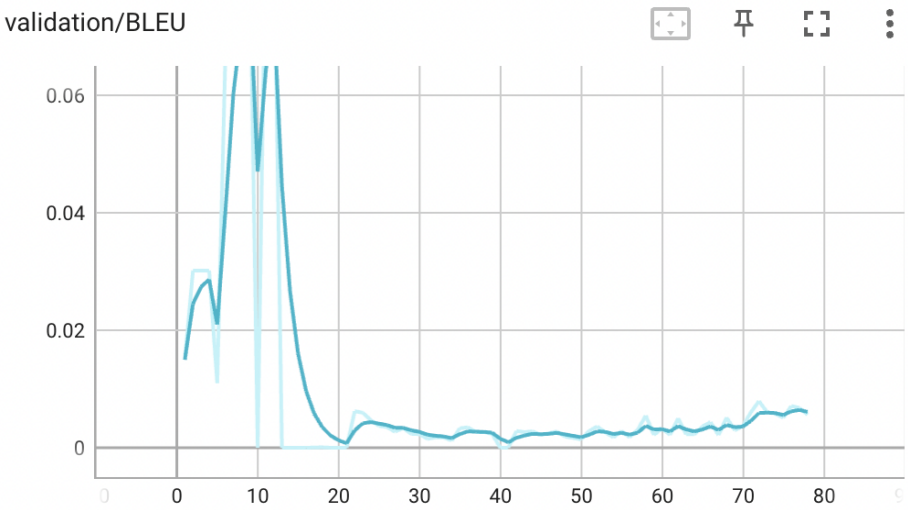
More sophisticated approaches, like including pairs of question and extra positives to augment data, or including information from a source link are left to the future.



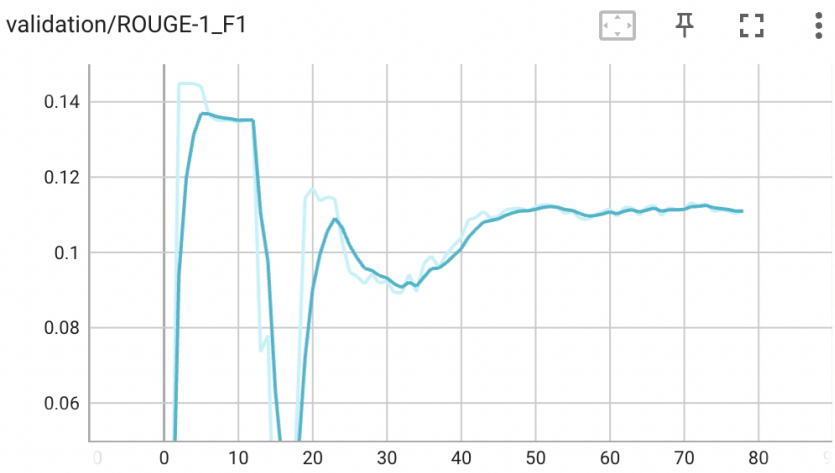
Train Loss Curve



Validation *Loss Curve*



BLEU scores on validation split



ROUGE-1_F1 scores on validation split

Looking at the training and validation loss curves, we see that model is actually training and fine-tuning for Own Game dataset. I used two automatic measurements for automatic results evaluation: BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). For the best checkpoint, validation BLEU is 0.005, and ROUGE-L is 0.11, which are very low values. There are metric spikes in the beginning of training, but they are not stable. Around 20 epoch, there is a significant drop, and then stable learning pattern can be observed. To sum up, validation metrics have low values and unstable patterns.

So, I draw a hypothesis that only multilingual language model is not enough for answering Own Game questions.

4.2 Manual Results Analysis

I also conduct manual inspection of test predictions. The most important thing to mention that only 2 of 2560 (0.08%) test questions were answered correctly, measured by exact match (EM) metric. Moreover, correct answers are very generic: ‘a’, when question was asking about a letter, and ‘Russian’, when question mentioned a language. Thus, these answers can be simply guessed without any neural model.

Another common problem of Natural Language Generation tasks is repetitiveness. For 2560 test questions model generated 467 unique answers. For example, ‘Alexander I’, Russian emperor, is an answer for many questions about rulers (English, French, Russian), but also for questions about people in general (singers, poets). We see that model detects that man has to be predicted, but for some reason it output ‘Alexander I’ for various cases.

Outputting Russian emperor even for questions, implicitly mentioning France or England, shows that this model struggles from bias towards Russian history and culture. Probably, this problem could be tackled by augmenting training data with Jeopardy questions translated to Russian.

Predicting Russian emperor for all questions, where ruler or just a man is required, could be described by term ‘hallucinations’ (Marcus, 2020). Big Language Models have no way to restrain from an answer, they generate it even when model has not clue for the correct one. Hallucinations are important research topic, gaining a lot of attention last years (Shuster et al., 2021, Zhou et al., 2021).

Another point of awareness for neural language models is a lack of control and interpretability. User has almost no idea, what was the reason of outputting the specific answer. Moreover, adjusting a model behaviour is a hard task, often requiring full re-training on a new data.

Positive thing to mention is that model usually does not mix entity types. When question is about country or sport, then common answers include ‘Canada’ and ‘tennis’. Interestingly, that the model does not simply detect key words, like ‘sport’ or ‘sportive’. It generates answer ‘tennis’ even if word ‘game’ or ‘competition’ is mentioned in the question, so neural model could be useful to avoid handwritten rules for natural language understanding.

Finishing an analysis on optimistic notion, I would like to mention that fine-tuned model does not mix language of the output despite the multilingual nature of the pre-training process. Generated answers are always in Russian.

4.3 Next Experiments

Summarising potential improvements, I see the following options for the future research:

- ***Improving Dataset Quality***

I believe that human annotation of the dataset will help to remove remaining parsing and scraping errors. Improved quality of source data can lead to better training results. I also believe that merging two editions of the dataset will result in a more quality data.

- ***Data Augmentation***

I see several options for Data Augmentation:

  - First, extra positives could be considered as new question-answer pairs, so we could draw new quality examples solely from the collected meta information.
  - Second, it is possible to automatically translate Jeopardy questions to Russian and mix them with original Own Game questions. Potentially, it could help to balance dataset and mitigate bias towards Russian history and culture.

- **Adding Source Passages**

Probably, adding source data to question text could help language model to learn important signals from the broader context. However, source passages are not always available, and at inference they are not provided.

- **Increasing Model Size**

I conducted all the experiments with mT5-base (580 million parameters) as an underlying model, but larger models are available:

model	parameters
mT5-base	580 mln
mT5-Large	1.2 billion
mT5-XL	3.7 billion
mT5-XXL	13 billion

In general, increasing model size leads to better results, but I am not sure that for Own Game even mT5-XXL will be enough. I hypothesise that we need combination of LM and external knowledge for this specific case.

5. Conclusion

This project achieved two goals, stated in the introduction:

1. Dataset of Jeopardy-like questions in Russian was collected.
2. Multilingual neural language model was fine-tuned to generate answers, given question text.

However, I can conclude that only neural language model is not enough for this task. It performs quite well in terms of Natural Language Understanding, but knowledge-intensive tasks like Own Game questions require external knowledge component. Thus, there is an obvious need to improve neural language model by some kind of structural data, like knowledge graphs or database of Wikipedia passages.

References

1. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine (Dunn et al., arxiv 2017)
2. Few-shot Learning with Retrieval Augmented Language Models (Izacard et al., arxiv 2022)
3. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension (Joshi et al., ACL 2017)
4. Natural Questions: A Benchmark for Question Answering Research (Kwiatkowski et al., TACL 2019)
5. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al., NeurIPS 2020)
6. ROUGE: A Package for Automatic Evaluation of Summaries (Lin, Chin-Yew, ACL 2004)
7. The next decade in ai: four steps towards robust artificial intelligence. (Marcus, Gary, arxiv 2020)
8. "A Russian Jeopardy! Data Set for Question-Answering Systems." (Mikhalkova, Elena, LREC 2022)
9. Bleu: a Method for Automatic Evaluation of Machine Translation (Papineni et al., ACL 2002)
10. Retrieval Augmentation Reduces Hallucination in Conversation (Shuster et al., EMNLP 2021)
11. How to build an open-domain question answering system? (Weng, Lilian, Lil'Log 2020)
12. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)
13. Detecting Hallucinated Content in Conditional Neural Sequence Generation (Zhou et al., ACL 2021)