Thesis Proposal

Background

Nowadays, Open-Domain Question Answering research is mostly focusing on English-language applications and solutions. In my thesis, I want to explore multilingual question answering using a Russian Jeopardy-like dataset.

Though the dataset is published, it is not properly described. I collected the same data independently, and believe that it can be improved. For instance, we can add hard negatives and topic information to the training data. Speaking about models to solve this dataset, there is only one proposed base QA model with 2% accuracy. So, I am sure that we can beat the current SOTA.

Research Ideas

Apart from improving dataset quality and beating SOTA, I see two main research questions:

1. I assume that a combination of the language model and external knowledge index can help to build a better solution for this dataset. However, what is the best way to provide a knowledge component? Should we retrieve Wikipedia passages, guided only by question text, or pre-selection of passages, based on other features (question topic) that can lead to better results? I also believe that we can try exploiting knowledge graphs for this task. It is also interesting to research the best way of interaction between LM and knowledge index.
2. English-oriented QA has strong models and already built knowledge components. Can we automatically translate the Russian dataset to English and apply existing QA models? Are gains from better QA models outperforming translation errors?

Conclusion

To sum up, I am mostly interested in neural retrieval research, described as research question 1. Question 2 seems to be natural to pose, but I am not sure if it fits well into the same thesis.

Appendix:

Format of the data: questions are grouped into topics consisting of 5 questions somewhat related to the topic with increasing complexity (from 1 to 5). Question context is usually around 1-3 sentences, sometimes requiring multiple assumptions to arrive at the answer. Answers are usually short factoid statements (≤5 words). Additionally, questions are designed for different purposes, and there are separate challenges for children, students, and adults, so different datasets have different complexity in human perception. In my opinion, checking how human perceived complexity correlates with machine results is an additional research point out of the scope of this thesis proposal. Here is a sample source data (in Russian) - https://db.chgk.info/txt/eu15stsv.txt. Please, use automatic translation tools to get a general idea.