Prakapenka Thesis Proposal

Neural-Symbolic methods for answering Jeopardy-like questions in Russian

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 Jeopardy-like (Own Game) dataset in Russian.

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

Research Ideas

Apart from improving dataset quality, I want to focus on improving the SOTA model for answer Own Game questions in Russian. I propose to use neural-symbolic methods to solve the task:

1. **Symbolic Methods**

Let us start with symbolic methods. There is a monograph Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases by Weikum et al. (2021), which gives a comprehensive overview of knowledge bases (KB) creation and curation process. Additionally, it provides a lot of use cases, including QA, and further references.

Knowledge bases can provide a precise answer with high degree of provenance and interpretability. Typical pipeline looks like this:

Text Preprocessing → Extraction of Named Entities + Relations → Knowledge Graphs Lookup

However, knowledge bases may often miss a required fact to answer a question, or extraction of named entities and relations may fail in case of complex sentences.

2. **Neural Methods**

Several years ago FAIR researchers presented a paper, proving that big language models store world knowledge in the parameters to some extent (Petroni et al., 2019). Later, this research was extended by considering importance of context for answers factuality (Petroni et al., 2020). In parallel, authors from Google presented a paper about T5 seq2seq model used as a simple QA system without any external knowledge (Roberts et al., 2020).

Trying to apply above-mentioned methods to Own Game dataset, I have conducted the experiment trying to fine-tune mT5 using 20k question-answer pairs in Russian. Resulting quality is extremely low (0.1%, 2/2560), but it was expected, since this dataset is very hard to be solved solely by neural seq2seq model.

I assume that combination of the language model and external knowledge index can help to build a better solution for this dataset. The idea behind Retrieval Augmented Generation (RAG) is to get relevant Wikipedia passages using neural retrieval techniques: DPR (Karpukhin et al., 2020) or, lately, Contriever (Izacard et al., 2022). RAG-based models (Lewis et al, 2019, Izacard et al., 2022) show prominent results for QA datasets in English.

3. **Neural-Symbolic Methods**

Both neural and symbolic methods have their own pros and cons.

To begin with, knowledge bases (KB) provide precise and factual answers. Source of the answer is often available, so it could be checked and corrected, if needed. However, KB do cover only limited (though huge) set of entities and relations, so predicting an answer about unknown fact is impossible. Moreover, applying KB for QA requires additional NLP steps (text preprocessing, named entity recognition and disambiguation), which implies additional source of possible errors.

On the other hand, neural model with retrieval augmented generation expects as an input only sentence in natural language, so no preprocessing is required. Generated answer is usually a grammatically correct phrase, but it could be factually wrong. This phenomena often referred as ‘hallucinations’ (Marcus, 2020). Additionally, It is hard to analyse why neural model outputted one specific answer, and not another. It is also hard to control the output, and inference can be slow.

Thus, I believe that we can combine these two approaches. I see the following pipeline:

1. Text Preprocessing + Named Entity Recognition + Entity Linking + Relation Extraction

2. Knowledge Base lookup. If it is successful, then we can output an answer and stop.
3. Try to predict an answer using retrieval augmented neural system. Then, based on answer likelihood, for example, we can decide whether to stop.
4. Combining neural model and KB. It is an open research question how to do it in the best way:
    a. For instance, neural systems are very good in context understanding due to training on large collection of texts. It could be useful for entity disambiguation and linking, for example.
    b. Additionally, neural output can be useful for narrowing the search over KB, since entity type is often right (human or sport), but entity itself could be wrong due to hallucinations.

Proposed neural-symbolic pipeline is a draft idea, so actual implementation can be different.

Conclusion

To sum up, I would like to explore neural-symbolic methods for answering Jeopardy-like questions in Russian. Smaller steps will include improving dataset quality, building independent symbolic and neural systems, and finding the best way to combine two approaches.

References

1. Few-shot Learning with Retrieval Augmented Language Models (Izacard et al., 2022)
2. Unsupervised dense information retrieval with contrastive learning (Izacard et al., 2022)
3. Dense Passage Retrieval for Open-Domain Question Answering (Karpukhin et al., EMNLP 2020)
4. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al., NeurIPS 2020)
5. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence (Marcus et al., 2020)
6. Language Models as Knowledge Bases? (Petroni et al., EMNLP 2019).

7. How Context Affects Language Models' Factual Predictions (Petroni et al., AKBC 2020)
8. How Much Knowledge Can You Pack Into the Parameters of a Language Model? (Roberts et al., EMNLP 2020)
9. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases (Weikum et al., 2021)

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).

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. I also provide an example question from my seminar project - https://github.com/tsimafeip/yoga-qa

Sample Entry

topic_name	question_value	question_text	answer
Океаны	*4*	*Океан, в отличие от своих братьев, не участвовал В НЕЙ, благодаря чему сохранил свое положение, а не был низвергнут в Тартар.*	*титаномахия*
Oceans	*4*	*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 of 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