YogaQA Project Proposal

Introduction

The major focus of this project is **open-domain QA for Russian language**.

First, I would like **to craft a new dataset** from a Russian analogue of Jeopardy called 'Своя игра' ('Svoya igra', 'Your Game'). 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. Here is a good example of sample topics, you can translate it to have an idea. Please note, that only topic and question-answer tuples are mandatory fields, so some data files do not contain source links, for example.

Second, I will **finetune mt5** as a baseline closed-book QA system. There is a sample notebook for finetuning t5 for answering questions in English, so I believe this seminar project is doable in 2 months.

These two steps are enough for the seminar project, but I would like to expand this work to a thesis. You can find my ideas later in this document.

Background

Data Format

Questions are grouped to the topics consisting of 5 questions-answer tuples somewhat related to the topic with an increasing complexity (from 1 to 5). Question text is usually around 1-3 sentences, sometimes requiring multiple assumptions to arrive to the answer. Answers are usually short factoid statements (≤5 words). Sometimes, link to the question source is available, as well as additional answers and answer-format guiding hints.

Human-Perceived Complexity

For example, if we have topic 'Germans', then asking about Richard Wagner is a good 1 question, and asking about Fritz Walter, a captain of German national football team who won 1954 FIFA World Cup, is a plausible factoid for 5 question. However, I believe that sometimes it could be easier to answer 5th question for a neural model, than 1st if the last one requires complex reasoning.

Additionally, questions are designed for different purposes. There are separate challenges for children, students and adults. Hence, different datasets have different complexity in human perception even beyond the topic range 1-5. In my opinion, checking how human perceived complexity correlates with machine results is an interesting research point.

Previous Work

I have checked the state of the QA in different languages on http://nlpprogress.com. There are some QA research only for 6 languages, including Russian. There is a single one existing QA dataset in Russian. It is SQuAD-like data, based on Wikipedia, so 'Svoya Igra' dataset is a novel approach with no previous work, to the best of my knowledge.

Experiments

Creating Dataset

I see two main sources of the data: public data available to everybody and private datasets from community.

Public Data

There is a public knowledge base of the questions: https://db.chgk.info/tour/SVOYAK. I have already scraped this web page. The format is unified, so it was easier to start. Initial version of the dataset consists of ~25k questions and presents a wide range of questions, created from 1996 to 2016 by various authors. I am sending the dataset as csv file along with this proposal.

Private Data

I have asked my friends to send me a bunch of raw data, so, currently, I have a lot of questions in different formats (txt, pdf, docx).

I would like to try unifying various documents to the single format. This data is more recent, and the important tendency of the last years is to write questions that require more indirect reasoning. So, 15-20 years ago authors were trying to ask complex factoid questions, like ('Name 7th highest mountain on the Earth'), but now focus is shifted towards easier facts, but more complex reasoning.

Nevertheless, both steps require substantial manual engineering, and private data curation is definitely out-of-the-scope of the seminar project.

Main Contributions

I believe that contributing this new dataset and training neural QA model on top it:

- expands QA research for non-English languages
- tests neural models in a new interesting setup, because questions of 'Svoya Igra' are often non-trivial in terms of required reasoning steps

I see the following future work:

- Applying more complex open-book approaches for QA, that utilize external knowledge (mDPR).
- We can also try translating dataset to English and applying English-language QA models.
- There are also a lot of video recordings of TV version of 'Svoya Igra', which opens a horizon for creating multimodal datasets.

Finally, I have a notion of training a model, that **jointly generates questions and answers them**. Topic+question_complexity setup allows us to set a proper external context. Based on a topic name, it is possible to retrieve relevant Wikipedia pages, for example. The last idea sound ambitious, but I am really excited about it.

Personal Background

I have worked as a Software Engineer (~4 years) with some experience in Python (~2 years). Apart from this, my previous Bachelor and Master's theses (both Computer Science) were about Machine Translation:

mostly new corpus crafting and applying JoeyNMT framework to train Transformer models. Additionally, I completed some NN-related courses and conducted several experiments using Keras, PyTorch and Tensorflow. Finally, I have worked with Prof. Koller on a Semantic Parsing task for 4 months (Python, Java, AllenNLP, Docker).