DeepQAMaker - A Data Driven Approach for Question-Answer Pair Generation for Movies and Novels

Thakar, Viral Bankimbhai
vthakar@torontomu.ca
501213983

Gole, Montgomery
mgole@torontomu.ca
501156495

February 2023

Summary

1 Problem Statement

For the advancement of research in fields like machine reading comprehension and question-answering, large-scale question-answer (QA) pairs are essential. Creating question-answer pairs from documents requires knowing how to pose a question and what the appropriate response is. Creation of such question-answer pairs is very labour intensive and error prone task. In this project we are interested to develop a data driven approach which can take a document as input and generates multiple question-answer pairs as output. To demonstrate our approach we will use scripts of movies and fictional novels to create question-answer pairs. Our approach is based on [1] and [2] papers.

2 Key Contributions

The key contributions in this project are:

- Data Collection
- Exploratory Data Analysis
- Model for Question-Answer Pairs Generation and Filteration
- Analysis through Qualitative Performance Metrics

3 Methodology

3.1 Data Collection

- Question-Answer-Context Data
  - The Stanford Question Answering Dataset (SQuaD) [4], [3] contains 50,000 unanswerable questions, along with 100,000 question-answer-context pairs.
- Movie Scripts and Novels
  - We will extract and curate movie scripts and novels using online data platforms such as kaggle and hugging face.
  - Here are some initial samples we extracted for this proposal:
    - Star Wars Movie Scripts
    - Fiction Stories

3.2 Exploratory Data Analysis

We will perform EDA by implementing following standard NLP EDA techniques.

- Word Tokenization
- Sentence Tokenization
- Histogram of Paragraph Length - Characters
- Histogram of Words per Paragraph
- Histogram of Avg Word Length per Paragraph
- Check and Plot Top K Stop Words
- Check and Plot Top K Non-Stop Words
- Create Bigrams and Plot Top K Bigrams
- Create Trigrams and Plot Top K Trigrams
- Topic Extraction and Prediction with LDA
- Named Entity Extraction

3.3 Question-Answer Generation and Filtration

Our Question-Answer Generation and Filteration workflow is based on WikiOmnia [1]

- The GPT-2 [2] version with 124 million parameters will be fine tuned with the SQuAD. Each training example will have three parts fed into the model: *[Context]*, *[Question]*, and *[Answer]* with *[End-of-line delimeter]*.
- We will use this fine-tuned model with our own *[Context]* extracted from Movie Scripts and Fictional Novels to extract question-answer pairs from them.
- We will use the pre-trained SQuAD BERT model to generate a "gold standard" answer for further evaluation in the Filtration step.
- There are 4 main criteria proposed in the WikiOmnia to filter out question-answer-context groups. An observation will be removed if:
  - More than 1 interrogative (who, what, which, whose, where, why, when, how) pronoun is present in a question.
  - 70% of the words in the generated answer are not matching the SQuAD BERT model's answer. Both answers are lemmatized before matching.

  - Named entities in question and/or answer are not present in the context.
  - A question-answer pair from the same context has a Levenshtein distance $\geq 70\%$ within the set created from the context.

4 Qualitative Performance Metrics

Each metric determined in this project will be compared to the same metric determined in the WikiOmnia [1].

- Question-Answer Diversity
  - The **self-BLEU** metric will be used to determine the diversity of the dataset created. This will be done by taking some $X\%$, where $X \in \mathrm{R}$, of the questions generated from the question-answer pairs, and finding the median BLEU score for each observation from the $X\%$. The lower the median self-BLEU, the more diverse the set is as BLEU calculates similarity of texts.
  - **Wh-word ratios** By first developing a baseline of a wh-question (who, what, which, where, why, when, how) count from the full corpus of SQuAD dataset questions, we will do the same with the questions we have generated in order to find the ratio of SQuAD to Filtered-Generated question-answer pairs.
- Question-Answer Quality
  - **Filter Percentage** Given the 4 step filtration methodology seen above, we will calculate the percentage of filtered Question-Answer-Context observations within the unfiltered Question-Answer-Context observations. This assumes that the filtration method acts as a test of quality for within the Question-Answer pairs.
  - **Automated Evaluation Pipeline** The Automated Evaluation section of the WikiOmnia [1] will not be replicated due to time constraints.

References

[1] Dina Pisarevskaya and Tatiana Shavrina. Wikiomnia: generative qa corpus on the whole russian wikipedia. *arXiv preprint arXiv:2204.08009*, 2022.

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[3] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.