CS 4740 Project 4 Proposal

Xinyun Tang (xt222), Zhuofan Li (zl329), Zhoutong Li (zl683)

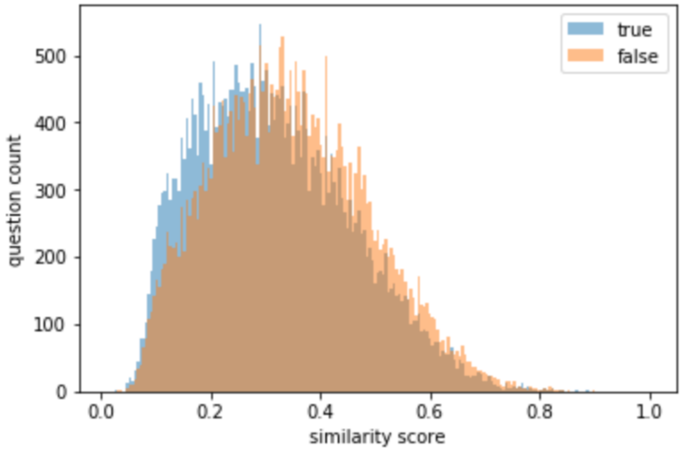
1. The Baseline System

    We implemented the baseline system to be a very simplistic information retrieval system with the context being the document we try to search and the questions being the query. More specifically, we used cosine similarity between the query vector and the document vector to decide how relevant the query is to the context. Then we took a look at the distribution of the similarity score to decide on a threshold to decide whether a question has a plausible answer.

    As for the tokenization method, since we are interested in the similarity between contents of the context and question, we used a TF-IDF tokenization approach where all English stop words were removed. Furthermore, the tokenization dictionary was built for each context independently.

    After tokenization, we calculated the cosine similarity score for all the questions and their corresponding context in training.json. The distribution for such similarity score for two categories were presented in Figure 1. As we can see from the plot, there is a large overlap between two categories and the distributions look very similar to one another. Though it may not be accurate, we still tried to set the threshold to be 0.25 and used the same similarity score on the development set. The precision and recall were XXX and XXX respectively.

    Figure 1. Histograms for the similarity score in two categories from training data



2. Proposed Final System

    As can be seen from the performance level from the baseline system, a simplistic IR system doesn't work very well for the current task. As a result, we made several plans for a more complex system.

    a. **Keyword extraction with coreference identification**

        As a first step, we extract some keywords from the question sentence using part of speech tagging and then try to find those keywords in the context with the help of some coreference extraction technique.

b. Concatenate the context and question tokens, together with its NER tag sequence and feed the sequence to a LSTM or GRU model with its “is_impossible” label as the final output.