Research Plan

Automatic question-answering through machine learning algorithms has numerous applications in business, public health, and other real-world problems where skilled work has the potential to be supplemented or replaced with AI natural language understanding. This investigation examined the AI2 Reasoning Challenge (ARC) question dataset, which provides multiple-choice questions to compare standard information retrieval methods that use text span with a more sophisticated two-step information retrieval protocol. The hypothesis is that this method will use contextual information to more accurately answer questions. The goal of artificial intelligence (AI) models that can process information from a large background text corpus is to produce useful knowledge that has the potential to supplement and even replace many functions that are now performed by a human being at great expense and after many years of training. The most ambitious of these goals is to eventually build systems that can imitate the basic abilities of skilled professionals in areas of the world too impoverished to afford the high cost of training and paying for services.

A Python script was written to process the training data, test data, and 14-millionsentence scientific fact database. Before applying either the experimental or control algorithm, the data will be cleaned of abnormalities and the corpus was preprocessed to filter out stop words, common words such as “the,” “an,” and “in” that do not provide meaningful data.

For the control group, a generic information retrieval model will be designed that uses a single step to identify key terms in the question and look them up in the corpus; this is based on the control model used by Ni et al., 2018 to compare with their distraction-removal algorithm. The control group’s algorithm selected its responses by identifying sentences in the corpus that shared words (stop words excluded) with the question. Using this subset of the sentences, the answer choice which had words in common with the most sentences will be selected as the prediction.

The experimental model will use two-step contextualization but standard protocol otherwise, keeping other variables constant. Whereas the control group will simply count the instances of co-occurrence between the question and answer text spans, the experimental group’s

algorithm will go a step further by analyzing other nearby words in those sentences. It will develop a score for each answer choice by adding the number of sentences sharing words with both the question and the answer choice to the number of sentences containing words in the context of those overlapping sentences.

Results will measure the accuracy using ARC's metric, by which a model can select as many answer choices as it wants for each question, and if any of them is correct the score is increased by a value of 1 divided by the number of answer choices that were selected. In addition to comparing against the control model, performance results will also be compared to existing models, the best of which are featured in the ARC website. The top model listed performs at 67.07% accuracy and the 10^{th}-best at 36.6%. The training time and run time of both models will also be recorded and compared with each other.

No changes were made to the research plan.