26 April
QANTA Entanglement
First Deliverable Update

I am writing this just as our Codalab manager confirms a successful run (as opposed to null running that we got earlier). Our paraphrase exploration has undergone some procedural changes, but remains the same in spirit–we are altering the way the guesser model interprets certain phrases via substituting paraphrases at different points of the training and guessing process and seeing how that affects accuracy.

Initial sifting through the original paraphrase database (PPDB) revealed that the highest quality, most 'accurate' paraphrases were often extremely colloquial (eg: what: what the hell), which is not relevant when taken in the context of Quizbowl question-sets. Fortunately, our part-of-speech manager found an offshoot of the original PPDB that tasked itself primarily with simplifying phrase. Many high-quality, simplified paraphrases take a superfluous and potentially confusing modifier (eg: extremely, very, much) and strip it from adjective phrases leaving a bare, 1gram adjective in place for vectorizing and processing.

Today's Codalab submission takes the smallest version of the above idea, with one bigram:1gram simplification (that has been confirmed to exist in the qanta devset) and runs it through our pipeline to make sure our overall replacement process is functional. During the generation of the QantaDataset, we take the incoming JSONs and do a regex check to see if their question text contains our bigram. If so, we are going to run it against a part-of-speech analyzer to make sure it is indeed the same phrase we are looking for and not part of an unrelated, but similarly formed phrase. This part-of-speech checker uses a process known as chunking, and is currently in progress. The matching phrases get swapped out of the JSON (courtesy of the team's data flow manager) that then gets trained.

This pipeline was more of a proof of concept. Our real aim is to incorporate these ideas 'live', as the guesser/scorer is taking in questions and determining the answers/probabilities. That will be our next step.

Additionally, our team will continue exploring the linguistic side of the problem, considering other possible relevant simplifications of phrases that could be confusing the guesser.