DEIIS Project Proposal
Group 5
Weston Feely, William Ibekwe, Siping Ji, Keerthiram Murugesan, Erzhuo Wang

Initial Pipeline/Workflow Design

Our baseline system will include two main modules: Annotation and Scoring. Annotation module focuses on deriving meaningful annotations from the input data using NLP utilities. Scoring module computes the aggregated score for each answer choice using the document and the background corpus. It does so by gathering a set of candidate sentences, for each question, collected from a simple synonym matcher and/or from querying SOLR index to support an answer choice. Some of the basic operations/methods in these modules are discussed in the later section of this document.

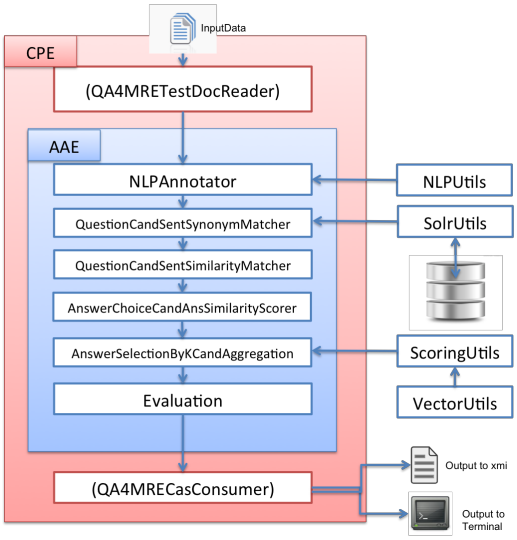

Figure 1: UIMA Workflow diagram for our Baseline System

Initial System Design

Our baseline simply consists of annotation of the answers, questions, and the mechanism of matching and selecting candidate questions/answers. The detailed description of the type systems can be found in the supplemental material TypeSystemDocumentation.txt.

Some improvement to the type system will include adding types to the current type systems that will help us improve the model. One Idea that we had in mind is creating a verb phrase annotation that would correspond to the noun phrase annotations that is in a baseline. We are also thinking about trimming down some systems to avoid redundancy of information. For example, some types like CandidateAnswer and CandidateQuestion both have the same content in other types. We will be investigating efficiency of these methods and other alternative methods to store similar data.

Baseline Methods and Improvements

Our improvements will include making a classifier for determining the type of each question, as described in the project guidelines. We plan on using every existing annotator in the baseline system, including the noun phrase annotator, named entity annotator, dependency

annotator, and synonym annotator. We will use these annotators and improve them to improve system performance.

For example, we will improve the noun phrase annotator by augmenting it with other parts-of-speech that are variants of the already included basic adjectives and nouns (JJS, JJR, NNS, NNP, NNPS in addition to NN and JJ). We will make a verb phrase annotator that will be similar to the noun phrase annotator in form, but will instead extract verb phrases using parts-of-speech like VB, VBZ, etc. We will improve the dependency annotator by filtering out dependencies that do not help the system, and we plan to do loose matching of terms using lemmas and synonyms. We will also improve the text segmenter by filtering more unnecessary segments from the data. Finally, we will implement the correct calculations of PMI and similarity scoring.

Division of Work

Wes and Keerthi will improve the existing annotators, starting with making such each annotator is working properly, and then improving each annotator's function as needed. Siping, Erzhuo, and Will will work on system improvements, starting with ensuring that the system scoring is working correctly. We will all post issues to github and work on the github wiki together.