Initial Write-Up

Team 9: Napat Luevisadpaibul, Dishan Gupta, Callie Vaughn, Wenyi Wang

Type System

Our types are as follows: SuperSentence, Sentence (extends SuperSentence), CandSentence (extends Sentence), Answer (extends SuperSentence),CandAnswer (extends Answer), Question (extends SuperSentence), Ngram,Unigram (extends Ngram), Bigram (extends Ngram), Trigram (extends Ngram), Document, SourceDocument (extends Document), TestDocument (extends Document), Phrase, Synonym (extends Phrase), NounPhrase (extends Phrase), NamedEntity (extends NounPhrase), Dependency.

Annotators

TextSegmenter, SentenceExtractor, NoiseFilter, NamedEntityRecognizer, PhraseExtractor, SynonymExpander, DependencyExtractor, NGramAnnotator, QuestionClassifier, AnswerChoiceAnnotator, QuestionAnnotator.

Pipeline

Our system will take in a background corpus and a test document. The background corpus will be read in by a background corpus reader and the test document will be read in by a test document reader. Both of these readers will output to our annotators, which will store their annotations in the CAS. The CAS will be input to the XMI CAS consumer, which will write out to the file system. The XMI background document reader and the XMI test document reader will read the XMI files from the file system. At this point the pipeline branches.

The XMI test document reader sends its output to a candidate sentence searching strategist, which then feeds the candidate sentences to the answer choice scoring strategist. At the same time, both XMI readers send their documents to the Solr/Lucene index, which searches the documents efficiently and sends its output to the candidate sentence searching strategist and the answer choice scoring strategist. Then the highest-scoring answer choice is selected by the answer choice selection strategist, which sends its selection to the result CAS consumer. This CAS consumer then writes its output to the file system.

Baseline Methods

We will first search for K relevant sentences using a combination of dependency, keyword and synonyms matching between the question and the sentence. For synonyms, we will try to expand the query directly for phrases instead of just individual words. Subsequently, the answer choices will be scored with respect to each relevant sentence based on noun phrase/name entity matching and synonym of noun phrase/name entity matching. Statistical measures such as Point-wise Mutual Information (PMI) and KL-Divergence will be used in scoring as well. Majority voting/aggregate score will be used to select the best answer finally. The CSE framework will aid in finding the optimal parameter configuration.

Division of Labor

In our group we all collaborated to come up with the pipeline, type system, and baseline methods. We then wrote the paper together.