Software Engineering Project Team 14

Initial Write Up

Team Member: Jiacong He, Qihui Li,
Haodong Liu, Qichen Pan

Git repository URL: https://github.com/table526/hw5-team14.git

1. Initial pipeline/workflow design

The basic structure of our system should have at least two main pipelines. One is responsible for preprocessing corpus files to build the original knowledge base of the whole system. The other pipeline serves as the runtime Q&A system to give best answers to a list of questions.

In preprocessing phase, our system will first use all annotators to annotate source documents to store the contents as CAS. Then system will index all the annotated documents using some tools such as Solr/Lucene.

In runtime Q&A phase, given a test document, our system will first find related sentences in knowledge base with the question. Then it will compare all given answers to the question with all the candidate sentences to select a best answer as the final output.

2. Initial type-system design

Document Level:

- SourceDocument: represents source documents.
- TestDocument: represents test documents that have questions and answers in them.

Sentence Level:

- Sentence: represents a single sentence.
- Question: represents a question sentence.
- Answer: represents an answer sentence.
- QuestionAnswerSet: stores all given answers of a question.
- CandidateSentence: represents sentence that may have relation with the question.
- CandidateAnswer: represents answer that may be the true answer (with PMI as parameter).

Analysis Element Level:

- Token: represents a single token in a sentence.
- Synonynm: represents words that may have same meaning as the given phrase.
- NounPhrase: represents a phrase that has a noun as its head word.
- NER: represent a token that is a named entity.
- Dependency: represent a relationship of dependency.

3. Baseline methods to implement & Task Division

The baseline system components and work division are shown below: