11-791 Homework 2

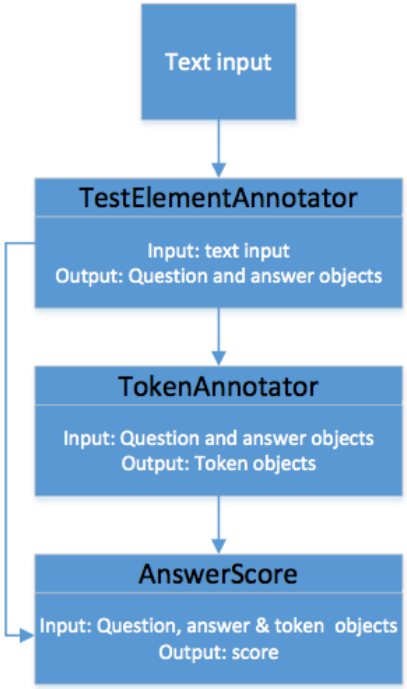
Wenyi Wang (Andrew ID: wenyiwan)

1. Introduction

Regarding the simple information processing task, my design is based on the hint in section 1.3 of the homework. My design aligns well with the goal of this homework, "simple".

2. Overall Structure

My design is based on the given type system. The "bag-of-word" theory works well.



3. Design Details

3.1 Test Element Annotation

TestElementAnnotator is in charge of annotating questions and answers. The input is the text file and output is question and answer objects (as defined in the type system).

3.2 Token Annotation

TokenAnnotator follows TestElementAnnotator and accepts question and answer objects as input. TokenAnnotator uses Stanford-coreNLP to tokenize sentences and output tokens.

3.3 Answer Scoring

AnswerScore accepts question and answer objects, as well as tokens and returns the score of each answer based on token (unigram) score.

3.4 Evaluation

This section returns the precision@N, where N is the amount of correct answers. The class "ScoreComparator" is a helper to achieve this.

4. Interesting Discovery

In the class "AnswerScore", I compared two ways of calculating score in order to find an effective and concise way of measuring the similarity between two bags of words. I compared (1) the cosine similarity of the two bags of words and (2) the number of overlapped words. The former seems more robust while the latter requires less computation. It turns out that their performances are the same given the same input.

5. Summary and Discussion

My idea on this homework is based on "bag-of-words". Although it is not perfect, it gives reasonable results. This simple implementation gives me a good image of how UIMA works and I really enjoy this framework. I hope to dig deeper and leverage this framework to build intelligent systems.