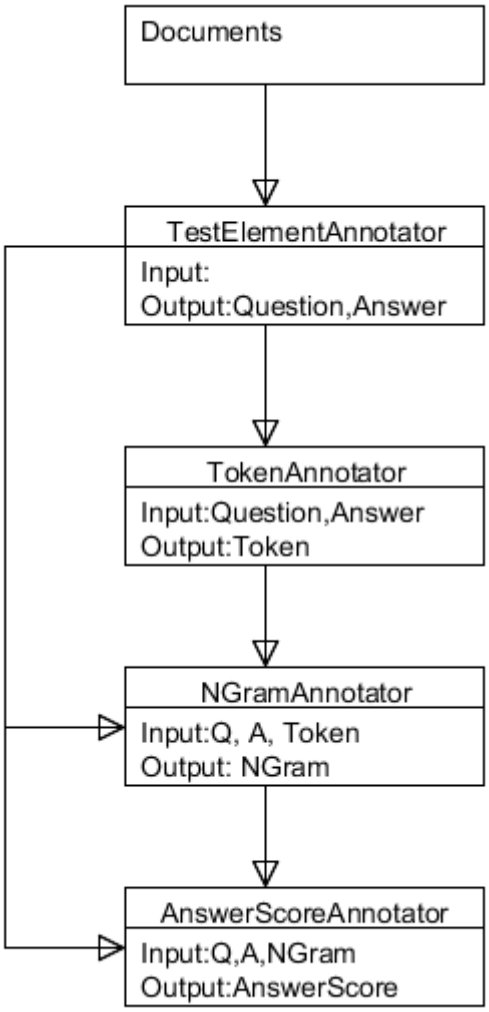
11-693 Software Methods for Biotechnology Homework2 Report

Shen Wu

1. System design:

In this homework, we suppose to implement the information processing task which we designed in the last homework. Type system is given so the focus in the implementation.
Basically, I followed the design of my last homework. TestElementAnnotator was created to annotate question and answers (O&As), and Q&As are outputted. TokenAnnotator take Q&As as input and further extract tokens. In TokenAnnotator StanfordNLP was used to tokenize sentences. NGramAnnotator take Q&As as well as Token as input because to extract NGram, it has to be known that each token from which sentences. Finally, AnswerScoreAnnotator gives the score of each answer based on NGram score.

2. UML description of flow-process diagram:



3. Interesting discovery and some detail

The hardest part for me is how to calculate the answer score of each answer. Because after tokens and Ngrams were extracted, we cannot know which question or answer is each token from based on given type system. And it's seems a waste of time re-tokenizing the sentence in the AnswerScoreAnnotator. My solution is to decide a token(Ngram)'s source by comparing its begin and end value with all the question and answers. In that way, the tokens and Ngrams do not need to be calculated again.

Also, to avoid looping of tokens, I use a hash table to store tokens. It's turns out to be much more efficient.

4. Comparison:

I implemented both Token Overlap and N-Gram Overlap. Turns out the later method performs much better because it considers the meaning of each word in the context of adjacent words. I wanted to improve the system through synonym search using WorldNet, but additional database is required and I was afraid this will be hard for evaluation. But I believe through searching synonym rather than simply comparing the strings, better performance can be achieved.