11693 Team 6
Final Project Report

Team members: Rui Wang, Victor Zhao,
Carol Cheng, Hua Tang, Yan Zhao

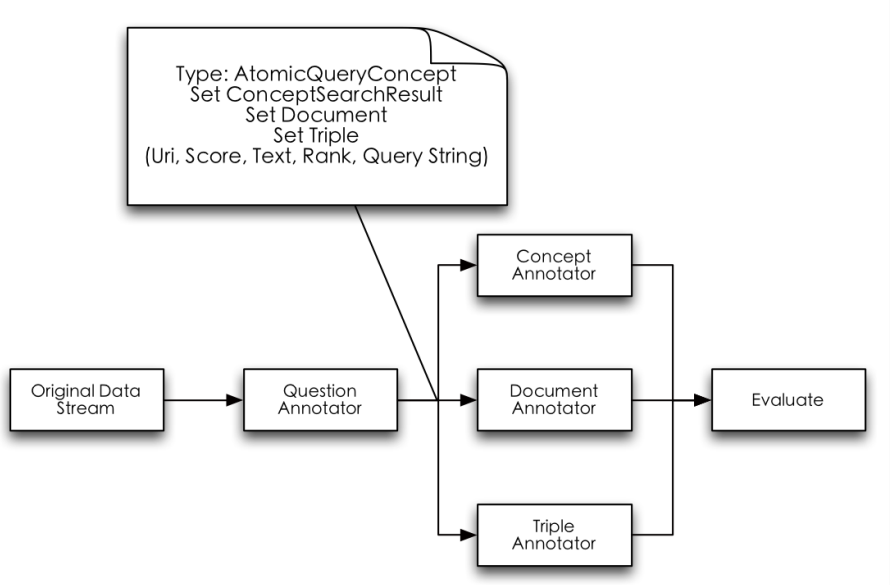
Milestone 1 Concepts, documents, and triples retrieval

Overview

In this milestone, our main task is to implement the simple UIMA pipeline and create annotators for concepts, documents and triples retrieval. We have to learn to use the utilities and web service API included in the archetype which can pass the query text to biomedical web services and retrieve concepts, documents and triples from these web services.

Architecture

The overall architecture for the current phase is quite straightforward. Apart from the general setting for the collection processing engine in UIMA, which consists of one collection reader, one aggregate analysis engine, and one cas consumer, a built-in type called AtomicQueryConcept has been adopted to store the content of queries passing to the following annotators. Moreover, the corresponding annotators store those processed results as features into this AtomicQueryConcept type.



Milestone 1 Architecture

Components

1. **CollectionReader**
   CollectionReader reads the input questions from a json file, stores the questions as Question type and put them into CAS.
2. **QuestionAnnotator**

Types:

a. Question
b. AtomicQueryConcept

QuestionAnnotator retrieves the questions from CAS, preprocesses the questions by removeing question marks, stop words and doing stemming on the tokens. QuestionAnnotator then stores the processed terms as AtomicQueryConcepts and put them into CAS.

3. **ConceptAnnotator**

Types:

a. Concept
b. ConceptSearchResult

This annotator uses OntologyServiceResponse service to process ComplexQueryConcepts and return several features which correspond to the data type, ConceptSearchResult, which was provided by project archetype. Finally,it stores the retrieved data into Concept type annotations.

4. **DocumentAnnotator**

Types:

a. Document
b. DocumentSearchResult

This annotator uses PubMedSearchServiceResponse service to process ComplexQueryConcepts and return several features which corresponds to the data type, DocumentSearchResult, which was provided by project archetype. Finally, it stores the retrieved data into Document type annotations.

5. **TripleAnnotator**

a. Triple
b. TripleSearchResult

The annotator uses LinkedLifeDataServiceResponse service to process ComplexQueryConcepts and return several features which corresponds to the data type, TripleSearchResult, which was provided by project archetype. Finally, it stores the retrieved data into Triple type annotations.

6. **CASConsumer**

Strategy

1. **Query preprocessing**
   a. **Punctuation removal**
      We remove the punctuations from the original questions. In this case, especially all of the question marks will be removed.
   b. **Tokenization**
      Tokenize the original questions by white-spaces.
   c. **Stemming**

We use the wrapper of Stanford stemmer provided by the homework 3 archetype to do stemming on the tokens.

d. **Stopword removal**

We remove words in question not affect the meaning of question, like “a”, “the” and “of”.

e. **AtomicQueryConcept**

After preprocessing the tokens, we save the tokens as AtomicQueryConcepts and put these annotations into CAS.

2. **Complex query**

a. **Operator**

When combing AtomicQueryConcepts, we can use operators like AND or OR. These operators are accepted by the provided web services.

b. **Structured query**

In order to get better results on concept retrieval, document retrieval and triple retrieval, we combine several AtomicQueryConcepts to ComplexQueryConcepts by using the operators like “AND” and “OR”. This method did help improve the overall performance of our biomedical question answering pipeline.

3. **Query expansion**

a. **Set score threshold for concept.**

In QuestionConceptAnnotator, we set score threshold for concept, which means

b. **Improve concept MAP**

c. **Extend query and retrieve documents**

4. **Document(Result orientation)**

Evaluation

So far, the usage of AtomicQueryConcept has meet the basic requirement of Milestone 1. Based on the current implementation, a couple of test sets have been tried and their performance results have been attached below.

The correctnesses on concept, doc, and triple have a great influence on the performance of snippetes retrieval and exact answer. So it is very important to evaluate the performance of this stage to think of an efficient way to get best performances on concept, doc, and triple.

For evaluation, we use unorder retrieval meatures including mean precision, recall, and F-measure and ordered retrieval meatures including MAP and GMAP to evaluate the results.

In the consumer, we store all the results of concept, doc and triple we retrieved. Then we build our metrics file in utils to help calculate the metrics, and then print our performance back in the consumer.

The overall performance of our system can be seen as followed. We get a high performance on concept based on the MAP metric, and the performance on doc and triple need to be further improved.

Metrics we used for single query:

1. Precision

2. Recall
3. F-measure

Metrics we used for whole documents of 29 queries:

1. MAP
2. GMAP

```
Query28:
precision:    1:  0.16666666666666666    2:  0.0    3:  0.0
recall:    1:  0.3333333333333333    2:  0.0    3:  0.0
fmeasure:    1:  0.2222222222222222    2:  0.0    3:  0.0

MAP:1:  0.4699233716475096      2:  0.0736659066663733  3:  0.0
GMAP:1:  0.057403421212056265   2:  0.002641663696531816        3:  0.001
Completed 29 documents
Total Time Elapsed: 3038103 ms
Initialization Time: 1875 ms
Processing Time: 3036228 ms
```

In the end, our team have successfully achieved of the minimum functionalities of the pipeline. In the next milestone, we may move on to the refinement of our query structure and the implementation of retrieval for snippets.

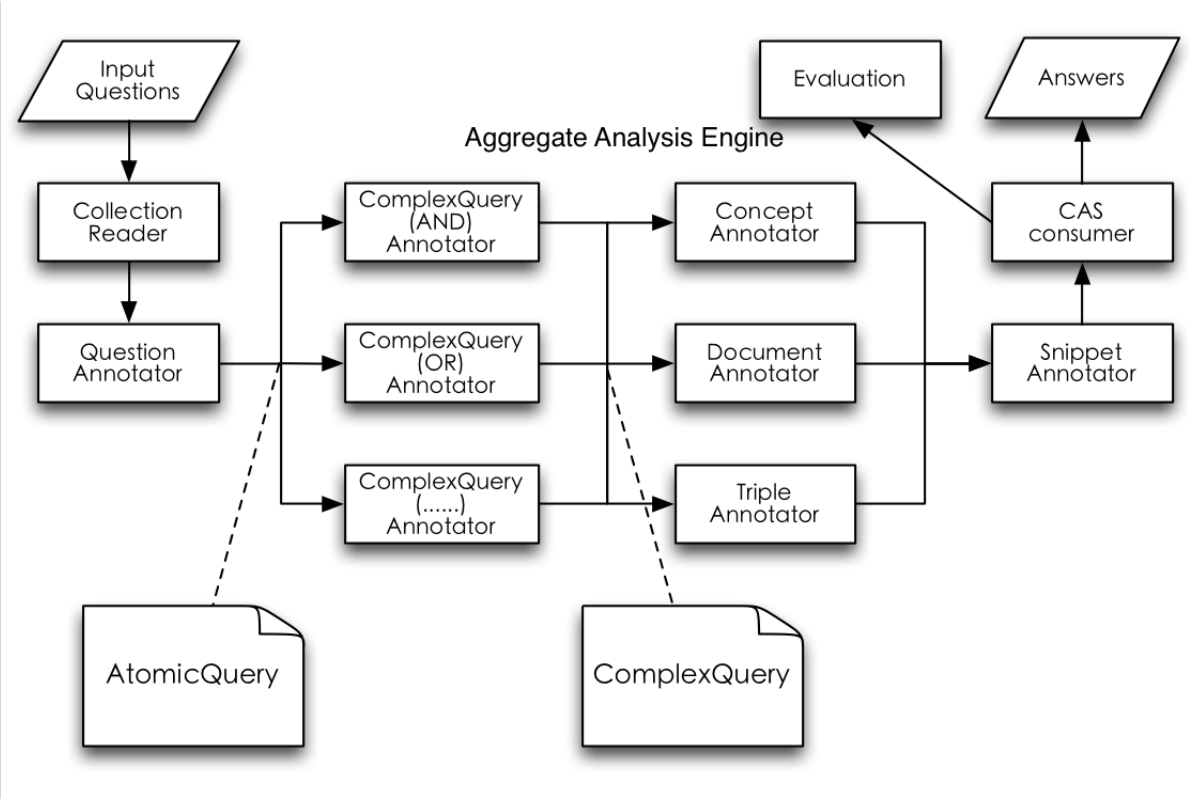
Milestone 2 Snippets retrieval

Overview

In this milestone, our task is to implement snippet retrieval component and integrate the component to our question answering system for a biomedical question answering pipeline. The snippet retrieval is based on the retrieved documents in milestone 1 and these snippets should be ordered in decreasing confidence. In addition, in this milestone, we also worked on improving question preprocessing including stop words removal, question mark removal and stemming.

Architecture

In milestone 2, we re-design our system architecture to the picture shown below:



Milestone 2 Architecture

New Components

1. **ComplexQueryAnnotator**
   ComplexQueryAnnotator combines several atomic query concepts to a white-space separated string.
2. **ComplexQueryANDAnnotator**
   ComplexQueryANDAnnotator combines several atomic query concepts by adding “AND” operater between them to produce a ComplexQueryConcepts.
   ConceptAnnotator, DocumentAnnotator and TripleAnnotator then use the ComplexQueryConcepts to query the web services and retrieve the concepts, documents and triples per question.
3. **SnippetAnnotator**
   Type:
   a. DocumentSearchResult
   b. Passage

SDSnippetAnnotator firstly gets the documents passed by SDDocumentAnnotator. Then, it calls the helper class, SimilarityCalculation, to do the similarity calculation and also get the article according to the given PMID. In order to make SnippetAnnotator work, we build two helper classes, SnippetRetrievalHelper, which allows our system to collect articles by the given documents. The function of another helper class, SimilarityCalculation, is to divide the article

into different snippets and compare them with the query string by calculating the similarity. In the end, this annotator will store the annotations into Passage type and save them to CAS.

Strategy

1. Document processing
2. Similarity
3. Information Retrieval

Evaluation

For evaluation, we maintain the metrics in the previous milestone, which use unordered retrieval measures including mean precision, recall, and F-measure and ordered retrieval measures including MAP and GMAP to evaluate the results of concept, doc and triple retrieval. Further, we use consine similarity between snippets and corresponding queries to evaluate the results of snippets retrieved, and for each article, we choose the sentence with highest similarity to add to the answer candidates. The performance can be seen as follow.

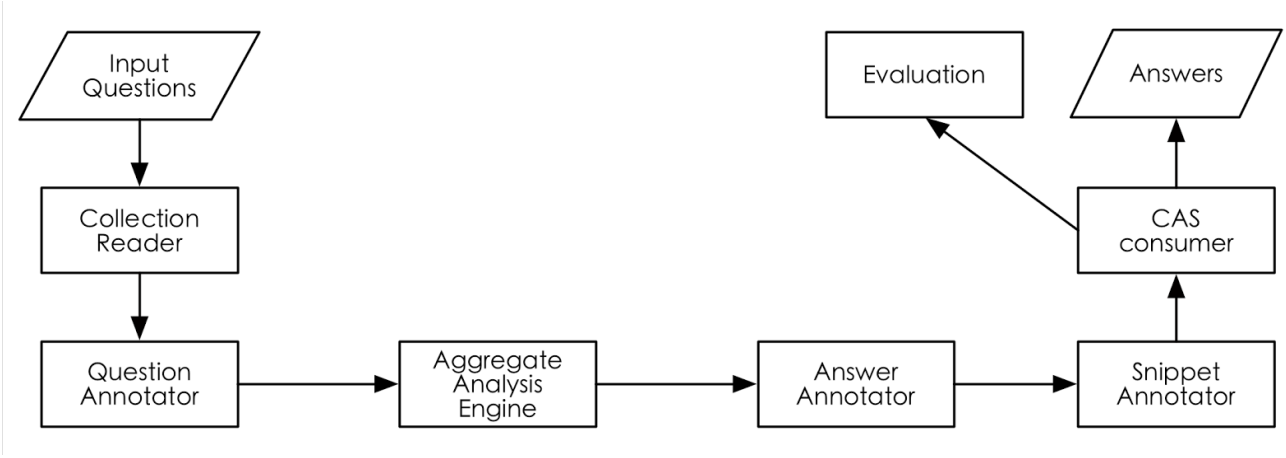
```
The 11's sentence is: This has been proposed to be responsible for improvement of certain autoimmune pathologies such as rheumatoid arthritis during pregnancy
@@@@@@ Similarity is:0.2236
VVVVVVVVVVVV
The stopArticle is : 1. Introduction Many connective tissue diseases have abnormal immune system activity with inflammation in tissues as a result of an immune system that
The query is:be rheumatoid arthritis more common in man or woman
The 12's sentence is: Similarly women have higher levels of prolactin and growth hormones compared to males
@@@@@@ Similarity is:0.0
VVVVVVVVVVVV
The stopArticle is : 1. Introduction Many connective tissue diseases have abnormal immune system activity with inflammation in tissues as a result of an immune system that
The query is:be rheumatoid arthritis more common in man or woman
The 13's sentence is: These pituitary hormones also enhance autoimmunity [1]
@@@@@@ Similarity is:0.0
VVVVVVVVVVVV
The stopArticle is : 1. Introduction Many connective tissue diseases have abnormal immune system activity with inflammation in tissues as a result of an immune system that
The query is:be rheumatoid arthritis more common in man or woman
The 14's sentence is: The effect of hormonal factor in inflammatory eye disease is complex and is not uniform among all diseases; male patients with Behçet's syndrome have
@@@@@@ Similarity is:0.0572
VVVVVVVVVVVV
The stopArticle is : 1. Introduction Many connective tissue diseases have abnormal immune system activity with inflammation in tissues as a result of an immune system that
The query is:be rheumatoid arthritis more common in man or woman
The 15's sentence is: Nonhormonal gender differences such as environmental exposures, drugs (more commonly used by one gender than the other), or infectious organisms can a
@@@@@@ Similarity is:0.063
VVVVVVVVVVVV
The stopArticle is : 1. Introduction Many connective tissue diseases have abnormal immune system activity with inflammation in tissues as a result of an immune system that
The query is:be rheumatoid arthritis more common in man or woman
The 16's sentence is: Autoimmune diseases affect the eye in different ways
@@@@@@ Similarity is:0.1111
VVVVVVVVVVVV
The stopArticle is : 1. Introduction Many connective tissue diseases have abnormal immune system activity with inflammation in tissues as a result of an immune system that
The query is:be rheumatoid arthritis more common in man or woman
The 17's sentence is: We will summarize the ocular manifestations of the most encountered connective tissue diseases and vasculitides
@@@@@@ Similarity is:0.0
maxId is11
```

Milestone 3 Exact answer generation

Overview

In this milestone, our main task is to fetch exact answr from snippets . We have to use different approaches to compare snippets with questions and try to find correct from candidates passages.

Architecture



New Components

1. AnswerGenAnnotator

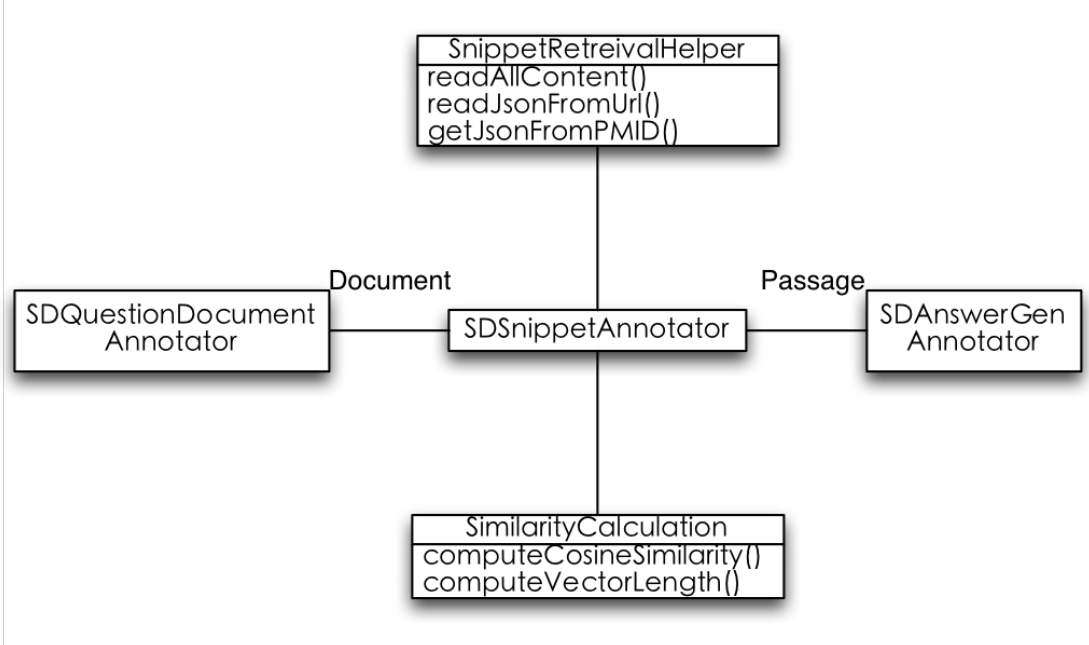
This annotator is in charging of producing exact answers. It fetches passage type from CAS and compute similarity between question and passages. After that, it will decide what answer it will give, “YES”, or “NO”.

a. Passage

This type is used for delivering snippets, where correct answers may be contained. however, because our system is yes/no system, we do not need to find answer but give yes/no based on snippets itself.

b. Answer

This type is used to save final answer to current processing question. It appears that if we find there is one satisfied similarity value.



Strategy

1. Biomedical name entity recognizer

This strategy is used for recognize Bio terms from query and passages. Based on these terms, we can compute similarity between term vectors. This is a very strong flag if we find high similarity between two victors. It suggests that “YES” answer has very high probability.

2. Threshold

Based on Vector Space Model, we utilize Jaccard-Dice similarity approach. But there should be threshold we need use if we want to convince us that we have can give “YES”. This threshold has been set as 0.1 based on experiments.

Performance

1,2,3 represents for concept, doc and triple

Without eliminating stopwords.

1. No concept score threshold

MAP:1: 0.4309209770846308 2: 0.11169953746658041 3: 0.0
GMAP:1: 0.12270265589772357 2: 0.0034635533317571043 3: 0.001
Completed 29 documents

2. Concept score > 0.1

MAP:1: 0.45114942528735635 2: 0.07434321562749996 3: 0.0
GMAP:1: 0.04628544171012127 2: 0.002342312669115188 3: 0.001
Completed 29 documents

3. Concept score > 0.15

MAP:1: 0.456896551724138 2: 0.07434321562749996 3: 0.0
GMAP:1: 0.03985055689816849 2: 0.002342312669115188 3: 0.001
Completed 29 documents

4. Concept score > 0.2

MAP:1: 0.3045977011494253 2: 0.07434321562749996 3: 0.0
GMAP:1: 0.012413856559730754 2: 0.002342312669115188 3: 0.001
Completed 29 documents

With eliminating stopwords.

1. No concept score threshold

MAP:1: 0.4309209770846308 2: 0.11169953746658041 3: 0.0
GMAP:1: 0.12270265589772357 2: 0.0034635533317571043 3: 0.001
Completed 29 documents

2. Concept score > 0.1

MAP:1: 0.5043202310897411 2: 0.11169953746658041 3: 0.0
GMAP:1: 0.08815374380371799 2: 0.0034635533317571043 3: 0.001
Completed 29 documents

3. Concept score > 0.15

MAP:1: 0.4846743295019158 2: 0.11169953746658041 3: 0.0
GMAP:1: 0.05022899948126503 2: 0.0034635533317571043 3: 0.001
Completed 29 documents

4. Concept score > 0.2

MAP:1: 0.45019157088122613 2: 0.11169953746658041 3: 0.0
GMAP:1: 0.03958141205023805 2: 0.0034635533317571043 3: 0.001
Completed 29 documents