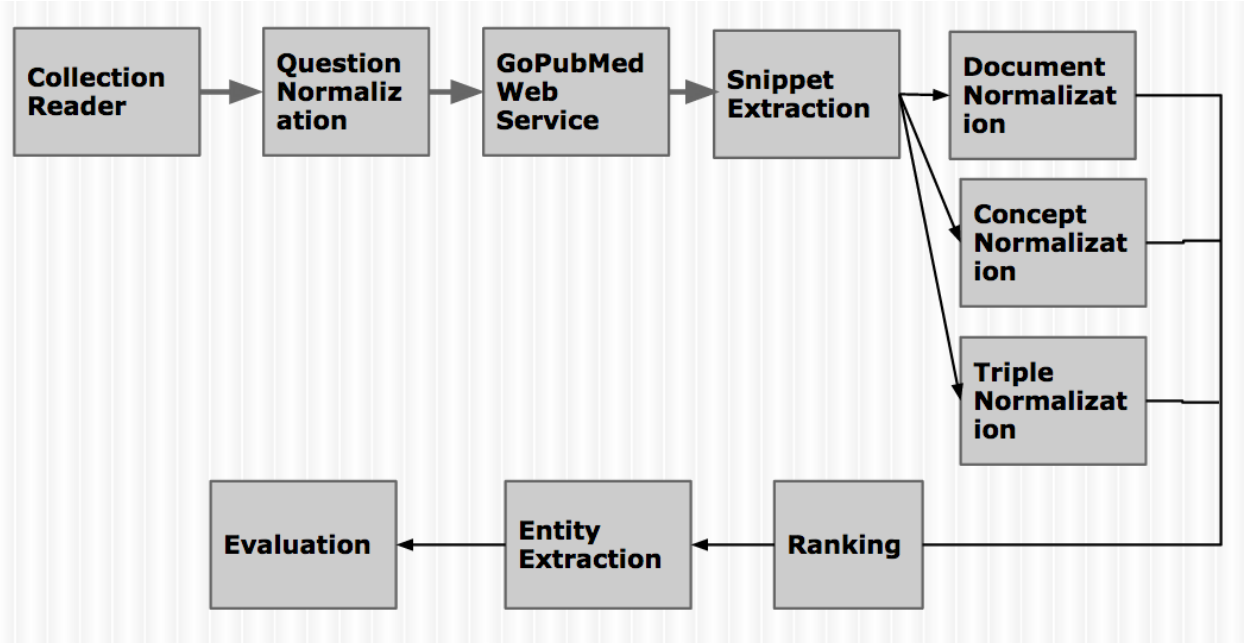
:

11791 Team 6 Final Report

Rahul Goutam, Yu-Hsin Kuo, Diyi Yang, Alok Kothari

Architecture



Collection Reader

Question Reader

The QuestionReader extends CollectionReader_ImplBase, which reads json formatted data from "/BioASQ-SampleData1B.json". The Question Reader outputs Question, Document, Concept, triples and snippets into the corresponding type systems. The Question is put into the type system Question Annotation.

How to Deal With Gold Standard

The question is put into the Question Annotation type. The gold standard documents, concepts, triples and snippets are marked with a special SearchID "__gold__" and are stored in UIMA types: Document, ConceptSearchResult, TripleSearchResult, and

Passage. Factoid returned gold standard answers are stored in Answer TypeSystem with a special rank "-11791".

In detail, we add several functions in util.TypeFacoty: createGoldStandardDocument, createGoldStandardConceptSearchResult, createGoldStandardTripleSearchResult, createGoldStandardPassage, and createGoldStandardAnswer. Then modify addQuestionToIndex in JsonReaderHelper to take into account gold standard cases.

Question Normalization

In this part, we discuss our approaches for query/question normalization. We list all the approaches used:

(1) Tokenization

(2) Stop words removal

(3) Punctuation/Number removal

(4) Stemming

(5) Using POS tag

(6) Query operation

Approach 1:

We combined (1) – (5) to do query normalization.

Approach 2:

However, when we do punctuation/number removal and stemming, some key words might be transformed into another word. For example, in this question, "Are there any DNMT3 proteins present in plants?" it transforms into "DNMT protein plant." We can see the DNMT3 becomes DNMT and this happens in other queries as well like Irg1 becomes Irg and Tpl2 becomes Tpl, etc. Therefore, we decided not to use number removal and stemming.

Approach 3:

Next, we applied Stanford POS tagger and then remove all the words that are verb in the query. The intuition behind this is that we believe as long as the nouns are present, it does not matter what the verbs are. Also, by reducing the number of the words in the query, we could also retrieve more documents than before and the result is below,

Approach 4:

At the end, we use simple “OR” operator to modify the original queries. However, it will returns too many irrelevant documents and degrades the overall performance. Thus we only use it when we cannot retrieve any documents for those queries.

GoPubMed Web Service

We have used GoPubMedService API to retrieve documents, concepts and triples. Documents are retrieved from the PubMed document collection. Triples are extracted from LinkedLifeData and concepts are retrieved from UniProtEntities, JochemEntities, GeneOntology, DiseaseOntology and MeshEntities.

The API takes the query as input and the number of documents/concepts/triples to be retrieved and queries the corresponding knowledge bases to get the results. The results are then put in the type system and added to the CAS. There were one problem with the API calls. For certain queries, the API threw an exception. There was no way for us to remove this exception because the error was thrown by the API which is 3rd party software. However, this would crash our pipeline. We handled this by catching the exception thrown but continued to run the pipeline for subsequent queries.

From the retrieved documents, we extract the text, title and URI. If full text of document is not available, we extract the abstract only.

Snippet Extraction

We assume that each snippet is a single sentence in the document text. We have used Stanford CoreNLP sentence splitter to split the text into sentences. Each sentence is then normalized and sent down the pipeline for ranking.

Normalization

The text of the document is normalized. Normalization of text has been discussed earlier.

Ranking:

The Ranking Part retrieves the scores for the Concepts and Triples. On documents and snippets, we assigns them scores using the Cosine Similarity function. (Snippets are passed on by the typesystem ‘Passage’.)

In order to do this, the snippets and documents are vectorized first by tokenizing , then creating term-frequency vectors and then doing L1 and L2 Normalization on those vectors. Given utility function: "rankedSearchResultsByScore" is then employed to obtain final rankings on documents, snippets, concepts and triples. These are then uploaded to the CAS.

Entity Ranking :

Entities need to be ranked according to relevance to question. Some techniques that we have used are :

* Term Frequency : Rank according to frequency of entity in retrieved documents. Higher the frequency, lower the rank.

* Tf-Idf : Rank according to tf-idf score of entity in retrieved documents.

* Tf-Idf Extended : Rank according to tf-idf score along with heuristics like length of term.

Entity Extraction

We extract entities for factoid questions from document text. We experimented with several techniques for extracting entities from document text. These techniques are :

* Named Entity Recognition : Named entities are extracted from text using HMMChunker from lingpipe. This model is trained on GeneTag data. It did not work very well.

* All words : We noticed that all factoid answers are single words. We can therefore extract each and every word and treat it as an entity.

Evaluation

The evaluator extends CasConsumer_ImplBase, which evaluates the results of our pipeline. The CAS stores both gold standard and retrieval results.

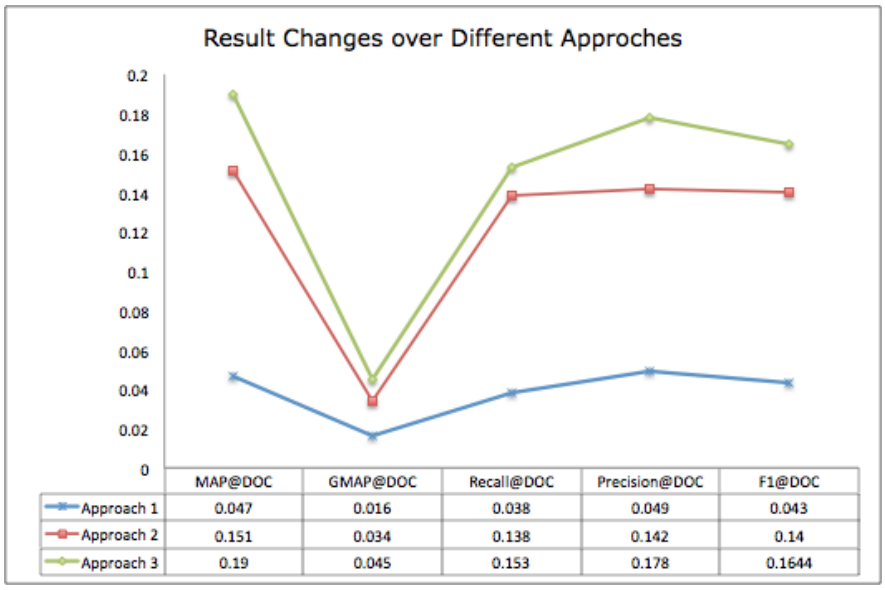
The Metrics that are used to measure the performance of document, concept, triples and snippets include Recall, Precision, F1, MAP and GAMP. For Factoid question, Strict and Lenient accuracy, MRR are calculated.

Results Comparison

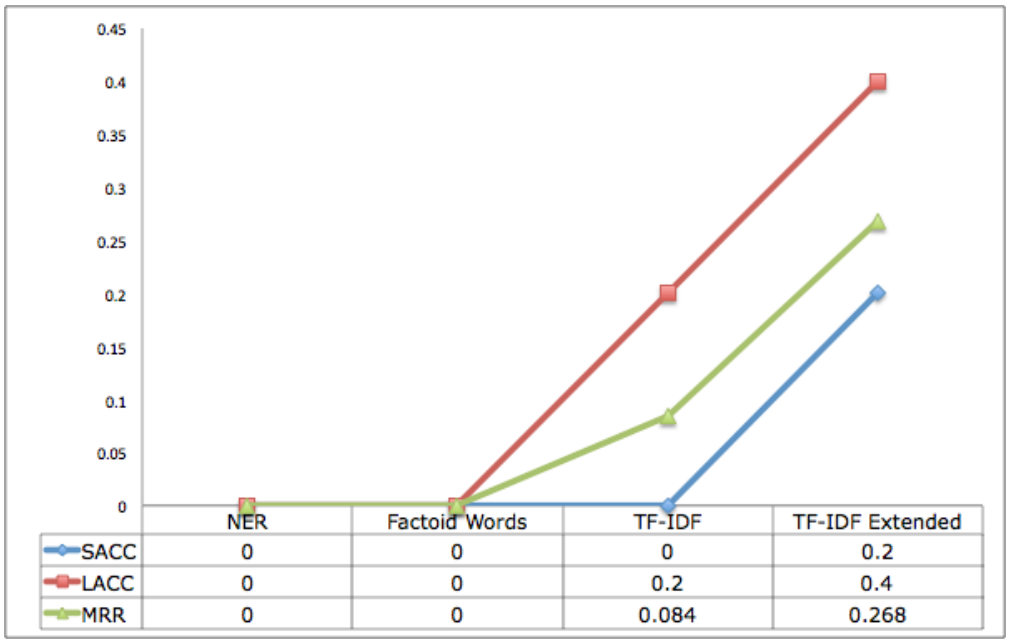
Retrieval Results

	Recall	Precision	F1	MAP	GMAP
Document	0.154	0.176	0.164	0.189	0.048
Concept					
Triples	0.110	8.357E-5	1.67E-4	0.004	0.012
Snippets	0.095	0.006	0.012	0.070	0.028

Retrieval Document Improvement



Factoid Comparison



Error Analysis

Rank 1	Rank 4	Rank 61
Original:Is Rheumatoid Arthritis more common in men or women? After processing: Rheumatoid Arthritis common men women	Original: Where in the cell do we find the protein Cep135? After processing: cell protein Cep135	Original: What is the methyl donor of DNA (cytosine-5)-methyltransferases? After processing: methyl donor DNA cytosine-5 methyltransferases
women rank : 1 men rank : 2 patients rank : 3 common rank : 4 rheumatoid rank : 5	cep135 rank : 1 protein rank : 2 centriole rank : 3 centrosome rank : 4 centrioles rank : 5	cep135 rank : 1 protein rank : 2 centriole rank : 3 centrosome rank : 4 centrioles rank : 5