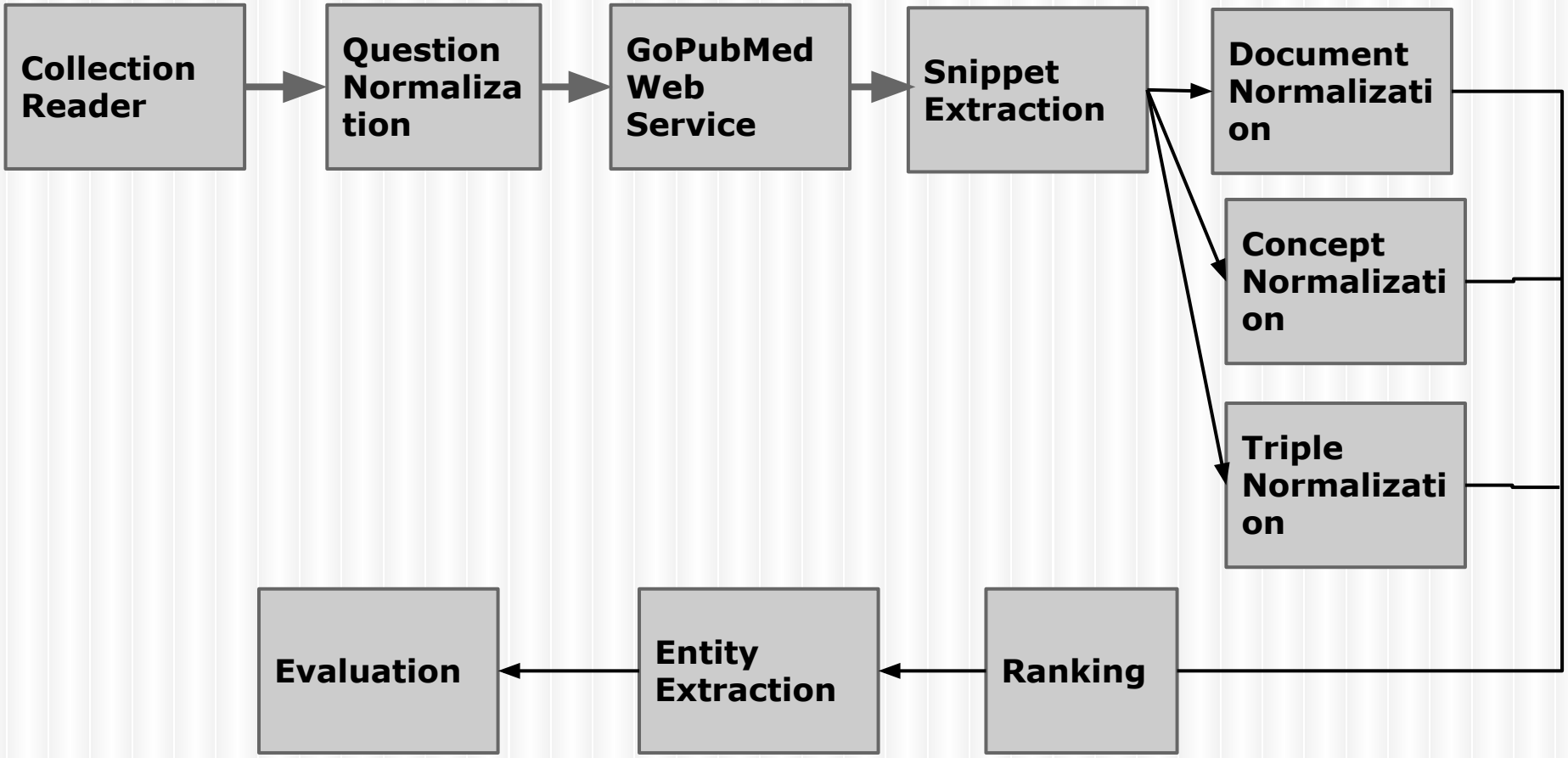
11791 Project

Team 06

Team Members

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

Architecture



Outline

- Question Reader
- Query Normalization
- Retrieval Analysis Engine
- Ranking
- Evaluation

Question Reader

- Collection Reader
  - Extends CollectionReader_ImplBase
  - Read Json file and store them into pipeline
  - Output question, document, concept, triples, snippets and answer (Factoid).
- Deal with Gold Standard
  - For concept, document, triples and snippets, a special searchId “__gold__”
  - For factoid answer, a special rank “-11791”

Question Normalization

(1) Tokenization

(2) Stop words removal

(3) Punctuation/Number removal

(4) Stemming

(5) Using POS tag

(6) Query reformulation

Question Normalization

- Approach 1: Combine (1) - (4)
- Problem

Are there any **DNMT3** proteins present in plants → **DNMT** protein plant

- Approach 2: Don’t use number/punctuation removal
- Approach 3: Remove verb words

GoPubMed Web Service

- API call used to retrieve documents, concepts, triples from KBs
- Input : Normalized query
- Refactor query if number of retrieved documents is less
- Exception handling

Snippet Extraction

- Snippet extracted from document text
- Decide on length of snippet
- Each sentence in text is a snippet
- Stanford CoreNLP sentence splitter used to split document text into sentences

Ranking Engine

- Retrieves the scores for Concepts and Triples.
- Ranks Document, Passage (Snippet) with Cosine Similarity.
- We do both L1 and L2 Normalization.
- Found Cosine Similarity to be an effective scoring function.
- rankedSearchResultsByScore is used for ranking.

Entity Extraction

- Entities extracted from documents for factoid questions only
- Named Entity Recognition
  - HMMChunker from lingpipe
    - Trained on Genetag data
- All words as entities
  - Factoid entities are single words
  - Use stanford corenlp for tokenization

Entity Ranking

- Entities ranked according to relevance of question
- Term Frequency - More frequent higher rank
- Tf-Idf score of entity in retrieved documents
- Tf-Idf with heuristics like length of entity name.

Evaluation

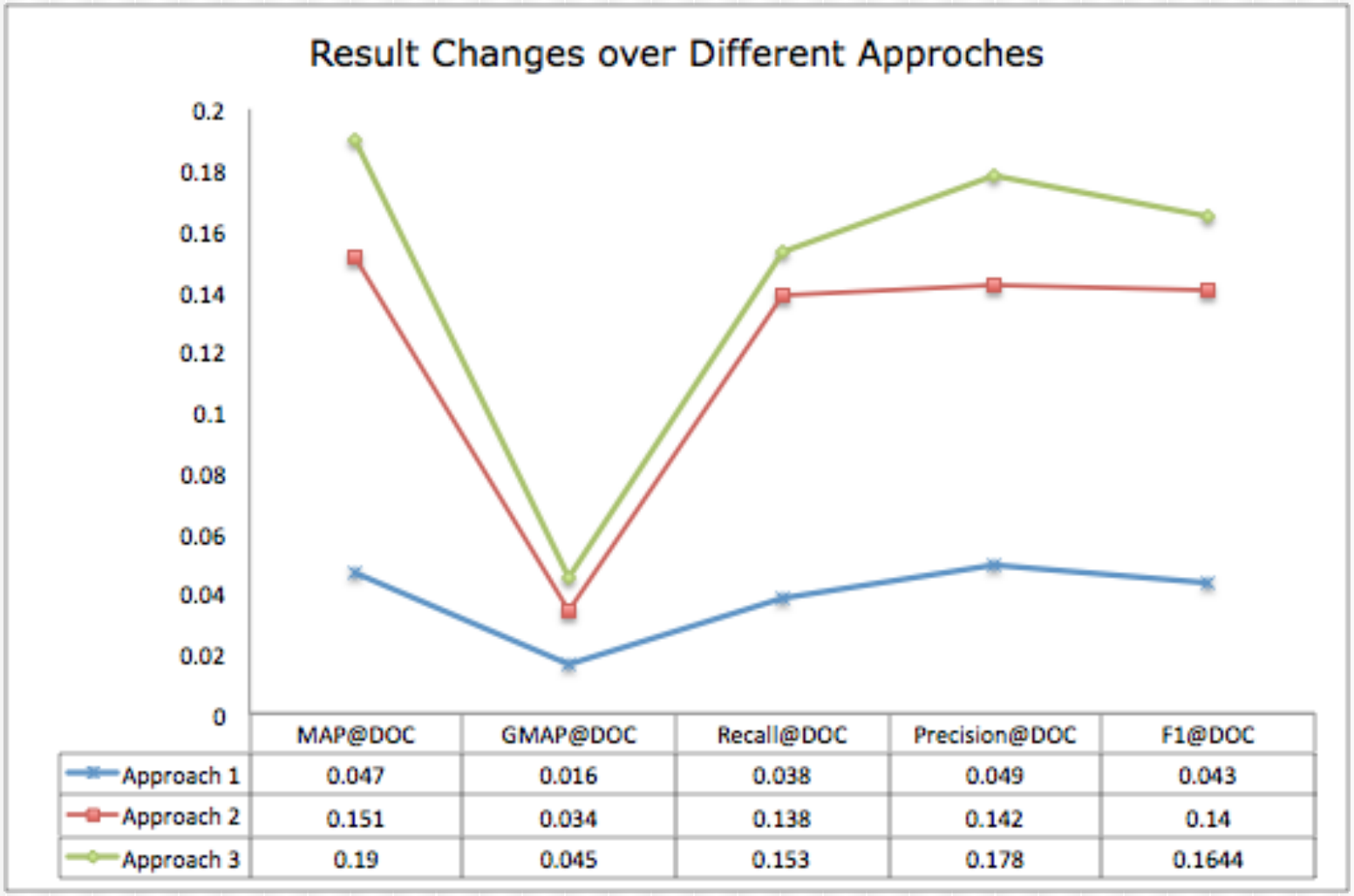
- Add Gold Standard during CollectionReader
  - addQuestionToIndex
- Evaluator extends CasConsumer_ImplBase
  - Evaluate the result against gold standard

- Evaluation Metrics
  - Unordered retrieval measures: mean precision, recall, F-measure
  - Ordered measures: GMAP, MAP
  - Factoid measures: Strict and Lenient Accuracy, MRR

Results for Retrieval
(Best Performance)

	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

Doc Retrieval Improvement



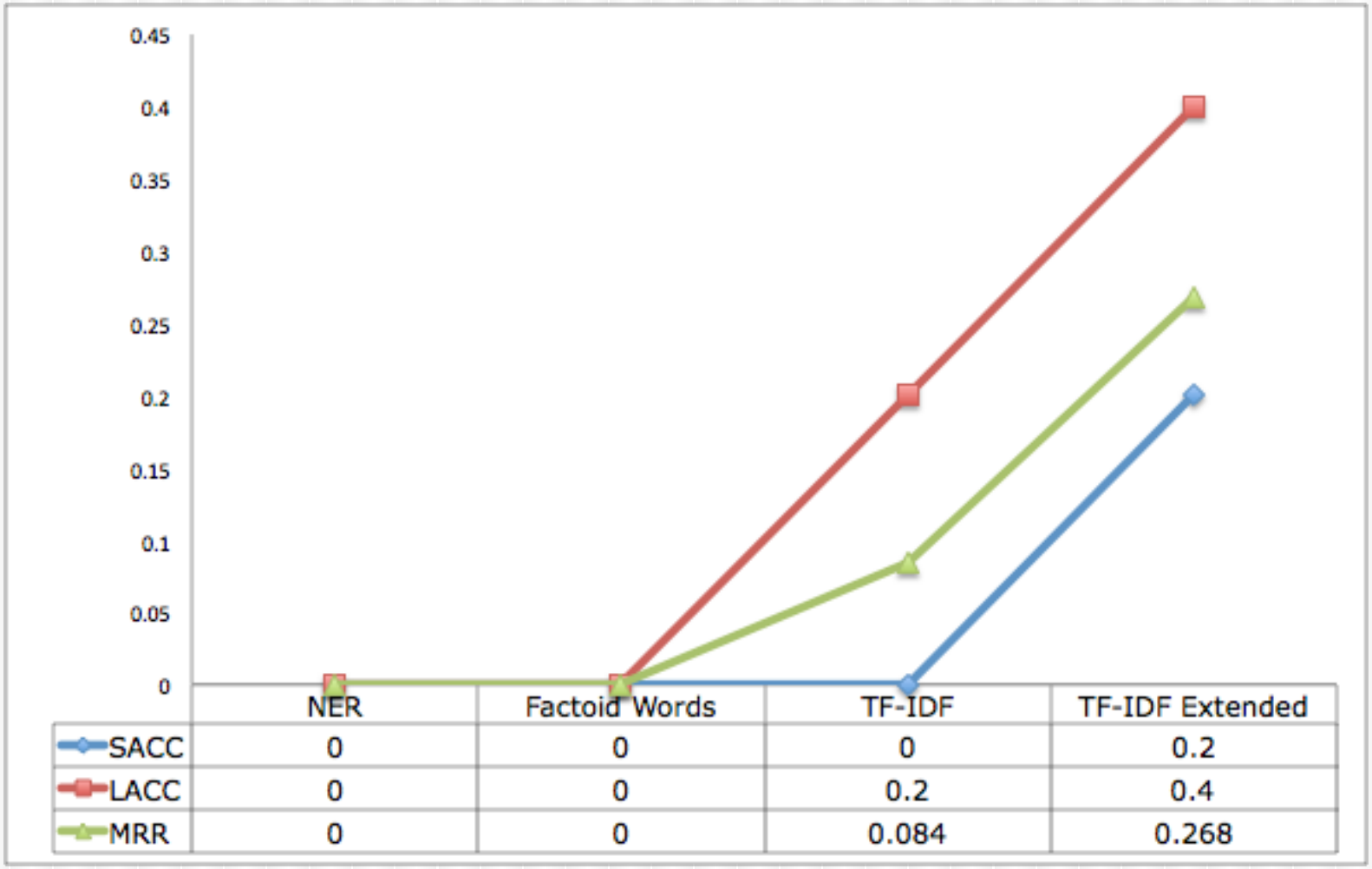
	MAP@DOC	GMAP@DOC	Recall@DOC	Precision@DOC	F1@DOC
Approach 1	0.047	0.016	0.038	0.049	0.043
Approach 2	0.151	0.034	0.138	0.142	0.14
Approach 3	0.19	0.045	0.153	0.178	0.1644

Results for Factoid

- Name Entity Recognizer (NER)
- Factoid Words
- TF-IDF
- TF-IDF Extended (removing short terms, length <= 2)

Approaches	SACC	LACC	MRR
NER	0	0	
Factoid Words	0	0	
TF-IDF	0	0.2	0.084
TF-IDF Extended	0.2	0.4	0.268

Results for Factoid



	NER	Factoid Words	TF-IDF	TF-IDF Extended
SACC	0	0	0	0.2
LACC	0	0	0.2	0.4
MRR	0	0	0.084	0.268

Error Analysis

- 5 factoid questions
- Individual rank: 1, 4, 20, 38, 61

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

Reflection

- Manage to work distantly
- Pair programming
- Communication effectively

Group Activities

- Eight Formal Meetings
- 30 Issues Assigned
- MANY informal talks via (Facebook vs. gChat)