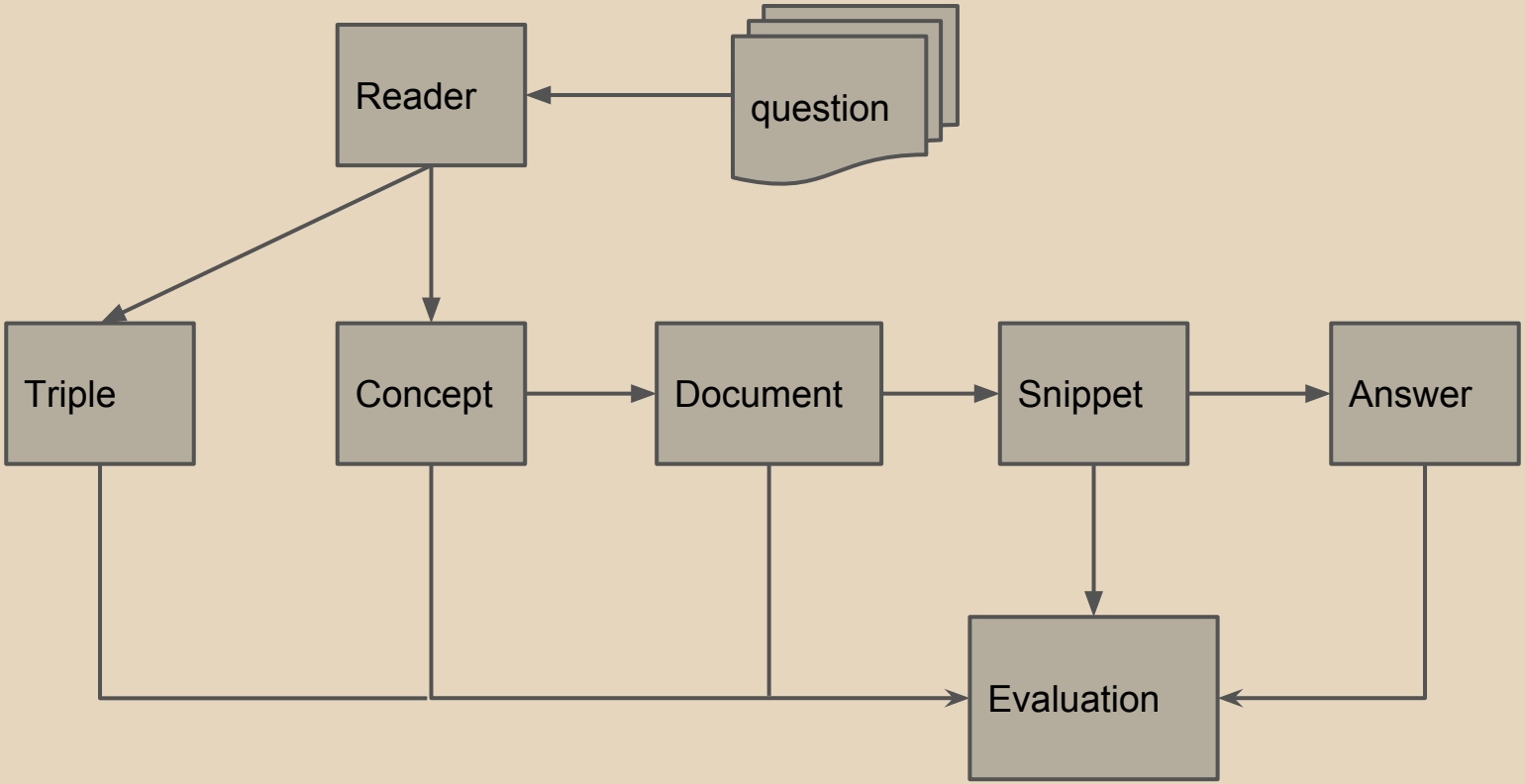
11-791 Project
Team 07 - Team HOLa

Hao Zhang
Oleg Iskra
Lara Martin

Outline

1. Pipeline Overview
2. Technical Details of AEs
3. Answer Types
4. Word2Vec
5. Exact answers
6. Team Tasks
7. Evaluation and Statistics

Pipeline Overview



Technical Details of AEs

Concept Analysis Engine
Document Analysis Engine
Snippet Analysis Engine
Answer Analysis Engine

Triple Analysis Engine

Concept Analysis Engine

Web Service Used:

DiseaseOntology, GeneOntology, JochemEntities, MeshEntities, UniprotEntities

Input to Web Service API:

QueryText.replaceAll("[?.,!:;]", " ")

Output we used:

finding.getMatchedLabel() finding.getConcept().getUri() finding.getScore()

Concept Analysis Engine

Ranking:

Using score we get from API

```
concept_hitsize = 20
TypeUtil.getScoredConceptSearchResults(jcas, concept_hitsize)
```

Document Analysis Engine

Web Service Used:

Pubmed

Input to Web Service API:

QueryText.replaceAll("[?.,!:;]", " ")

Named entities in query string

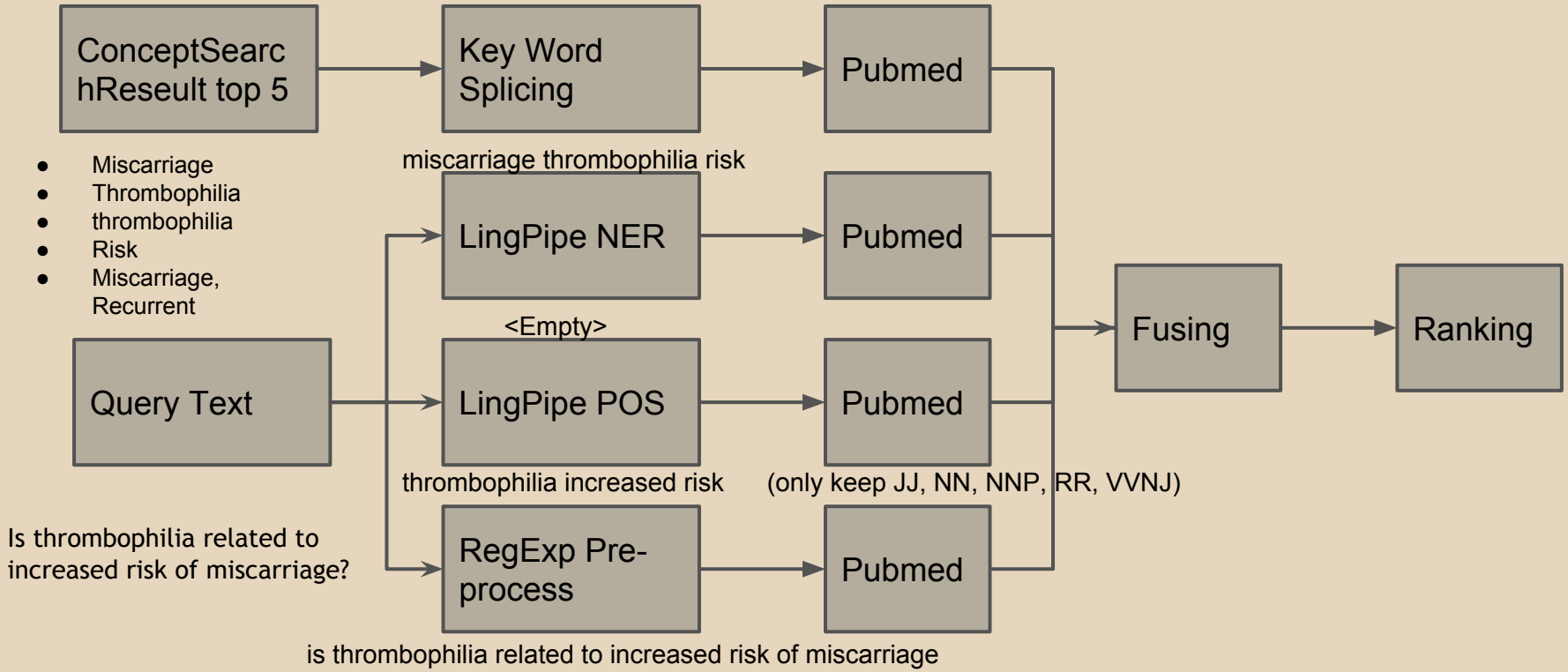
MatchedLabel from ConceptSearchResult

Query text filtering with part-of-speech tagger

Output we used:

finding.getPmid() finding.getDocumentsAbstract() finding.getTitle()

Document Analysis Engine



Document Analysis Engine

Ranking:

Score(D,Q) = α · f(D.abstract,Q.text)+(1-α) · f(D.title,Q.text)

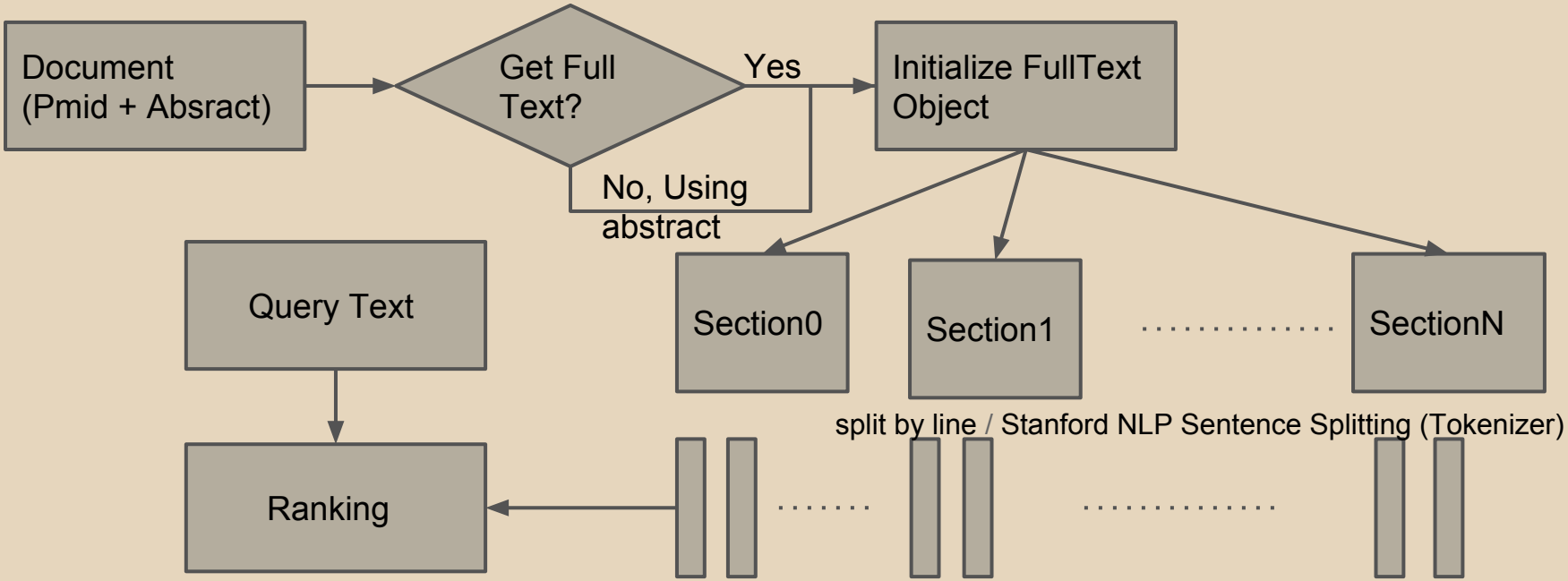
where:

f(x,y)=(<x,y>)/(|x||y|) / BM25 (hyper-params b, k1)

α = 0.5 (the value of α is **IMPORTANT**)

```
document_hitsize = 100
TypeUtil.getScoredDocument(jcas, document_hitsize)
```

Snippet Analysis Engine



Snippet Analysis Engine

Ranking:

Standard Cosine Distance

Score(S,Q) = f(S.text,Q.text) f(x,y)=(<x,y>)/(|x||y|)

Word Expansion with WordNet

Pseudo Code

```
for(String s : query)
    Set<String> s_synonyms = WordNetAPI(s);
    for(String t : sentence)
        Set<String> t_synonyms = WordNetAPI(t);
        if(s_synonyms ∩ t_synonyms)
            update(score); //Just using Cosine Distance
score = Normalize(score);
```

```
int snippet_hitsize = 50;    TypeUtil.getScoredPassage(jcas, snippet_hitsize);
```

Concepts improvement

Anticipated to use Continuous Space Word Vectors Obtained by Applying Word2Vec to Abstracts of Biomedical Articles from http://bioasq.lip6.fr/tools/BioASQword2vec/

Word2Vec statistics

During the European project BioASQ3, word2vec was applied to more than 10M English abstracts of biomedical articles from PubMed. The resulting vectors of 1,701,632 distinct words (types) are now publicly available at http://bioasq.lip6.fr/tools/BioASQword2vec/.

Word2Vec obstacles

- Main tool oficially available https://code.google.com/p/word2vec/ written in C++
- Similar framework in Java http://deeplearning4j.org/word2vec.html uses different file formats and different approaches and ND4J computational framework (http://nd4j.org)
- Hard time to integrate with existed UIMA pipeline and resolve Maven dependencies
- Hard time to solve all “Out of memory” and “Heap space” errors

Word2Vec quality

protein: proteins, a-anchoring, pka-anchoring

thyroid: thyroidal, nonthyroid, hyperfunctioning

associated: correlated, related, correlates

hormone: gh, luetinizing, fshluteinizing

human: murine, mouse, immortalized

used: utilized, employed, applied

Word2Vec reality (problems)

- Data file more than 3Gb needs more than 1 minute to load
- Uses more than 8Gb of RAM with the peaks up to 12Gb during loading
- Similar/related word search also relatively slow (approx. 800ms per query)
- Between many relevant items there are at least 20% completely irrelevant.

Exact answers

Identified type of questions:

- which, when, where, who, how, why, etc. - factoid questions
- am, is, are, do, does, did, shall, will, should, would, etc. - yes/no questions

Sources:

- Question Classification using Head Words and their Hypernyms, Zhiheng Huang, Marcus Thint, Zengchang Qin

Answering Yes/No and Factoid questions

- StanfordNLP and LingPipe annotators to break question into meaningful structure
- Use triples as key to question meaning (considering low triples quality)
- Analysis of related concepts and documents with related keywords to identify exact answer

Tasks

Hao - Answer Analysis Engine

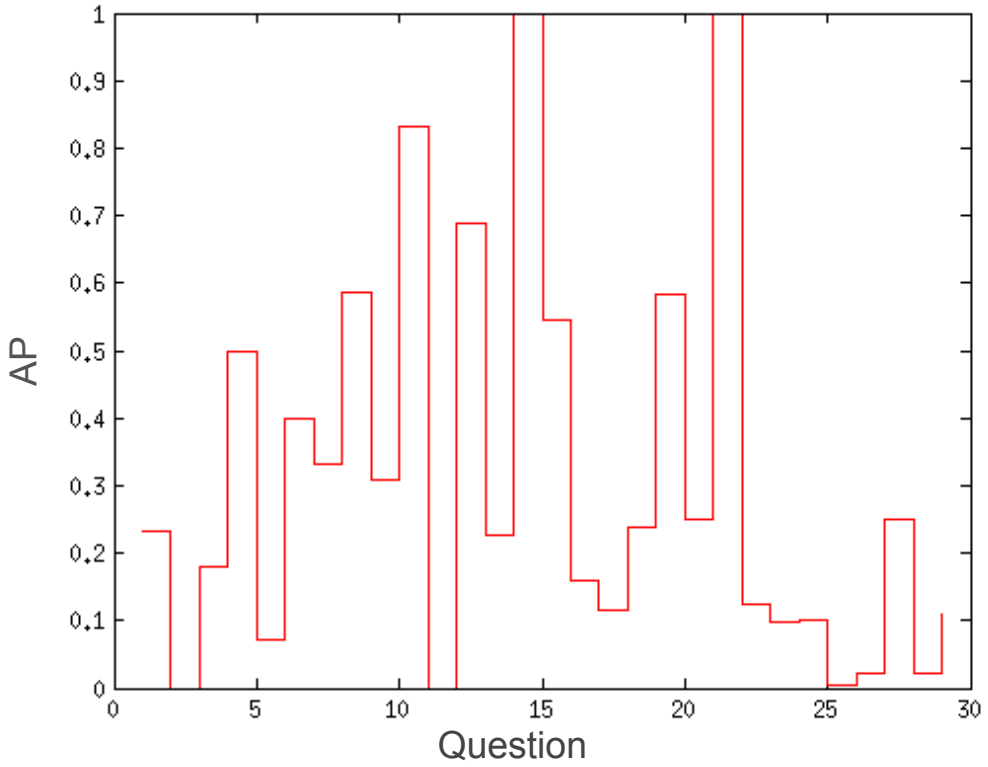
Oleg - Concept Retrieval Improvement

Lara - Evaluation

Evaluation

- Concepts, Documents
- Snippets
- MAP only; GMAP not helpful

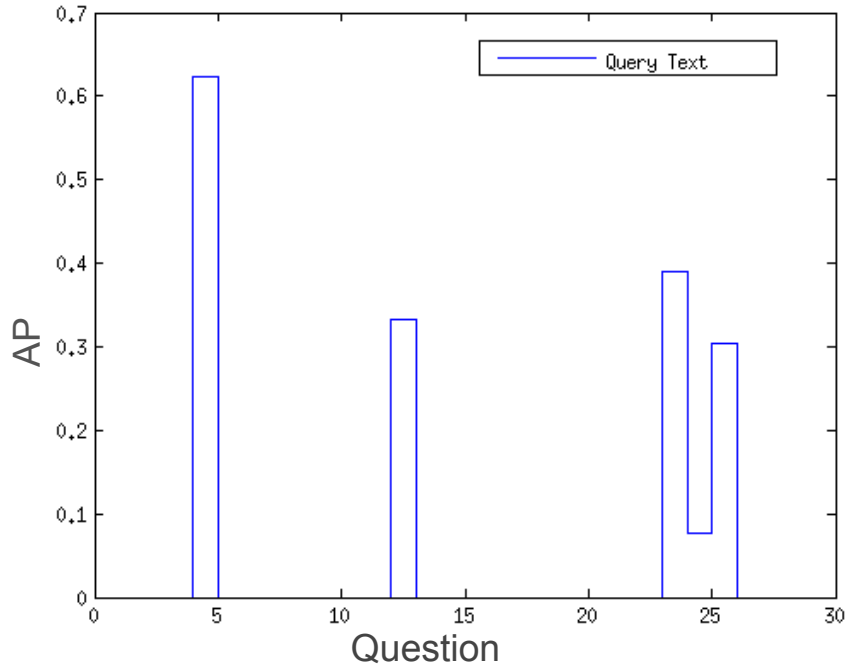
Statistics - concepts



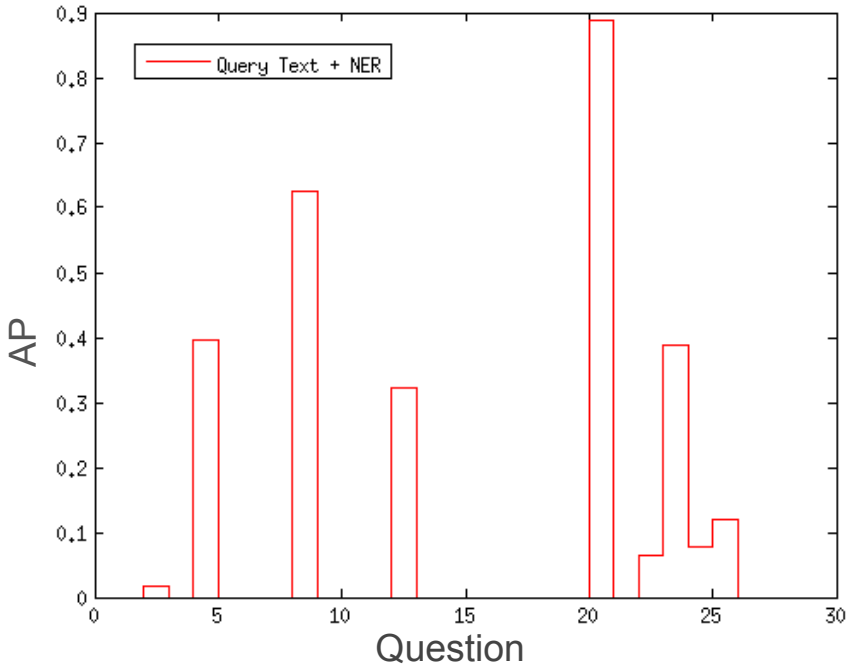
MAP: 0.3098

(from 0.1954)

Statistics - documents

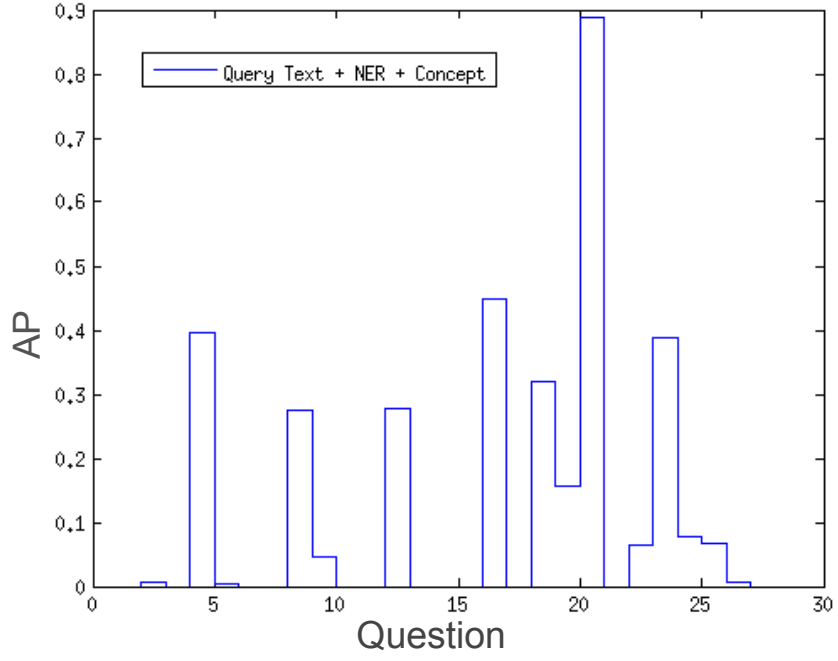


MAP: 0.0595

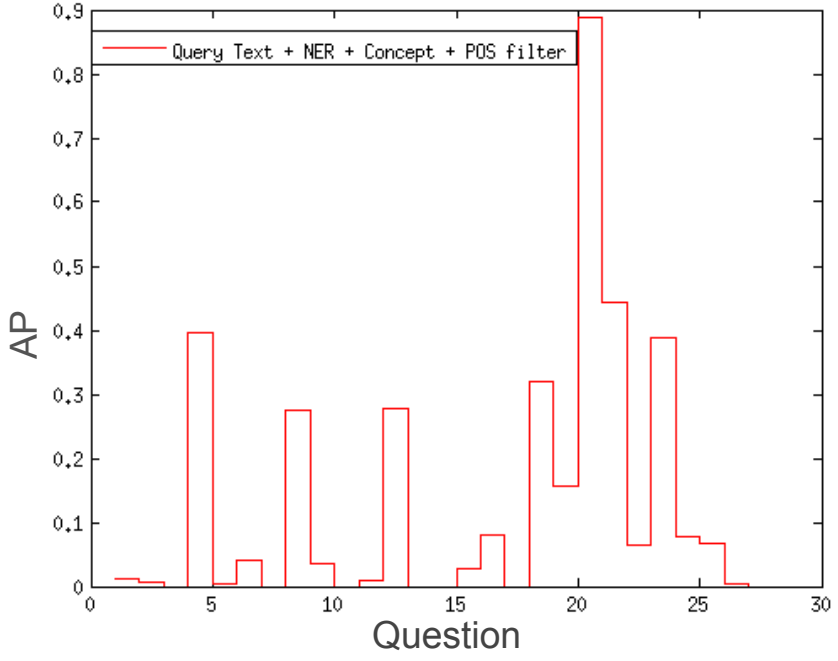


MAP: 0.0999

Statistics - documents



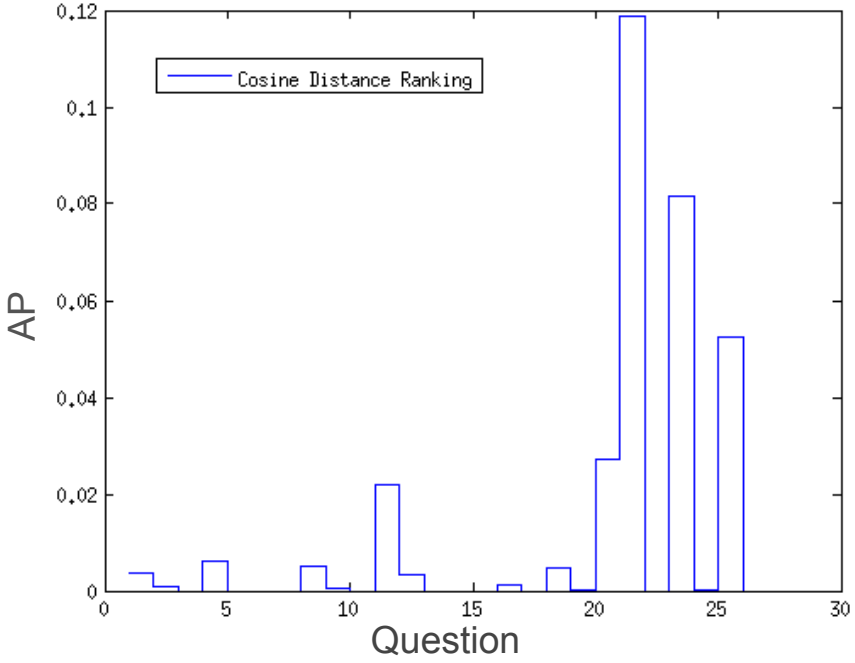
MAP: 0.1181



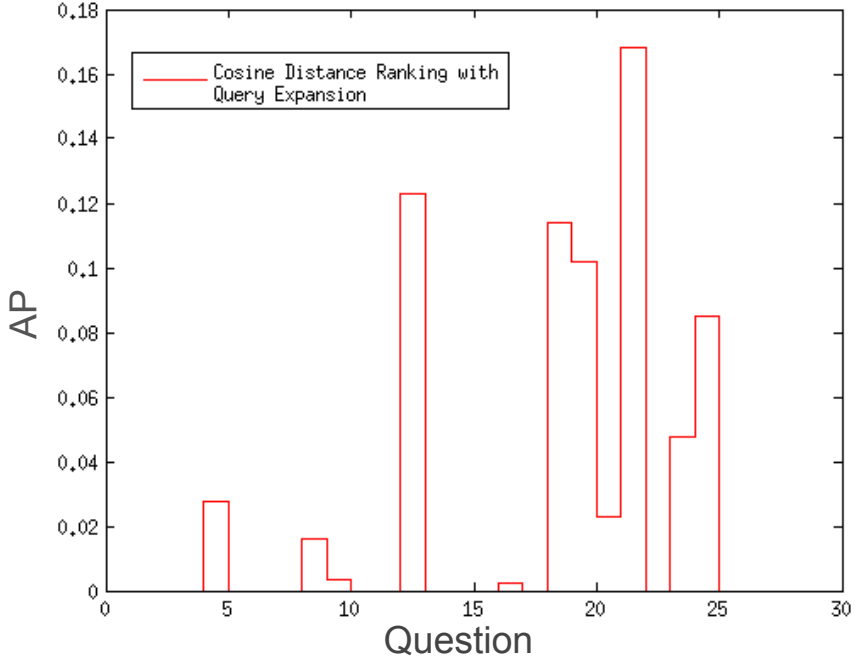
MAP: 0.1234

(from 0.0110)

Statistics - snippets



MAP: 0.0113



MAP: 0.0245

Thanks