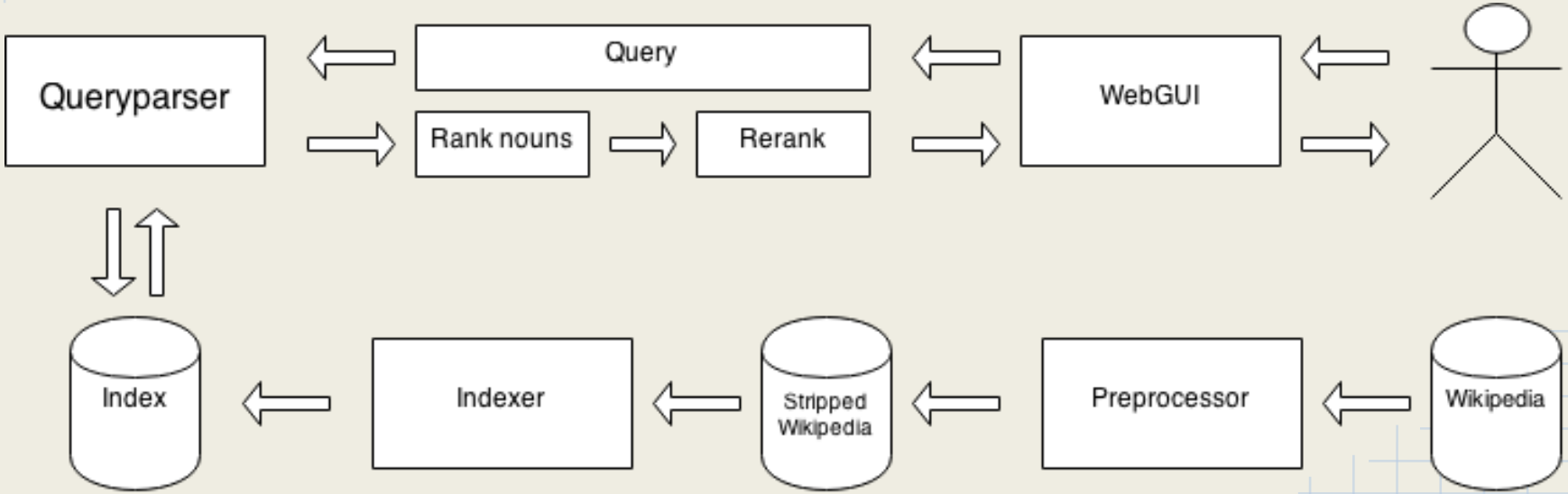
Minerva

Question-answering system

Background

- IBM Watson
- Wikipedia as the information source
- Parse natural language
- Given a question, how do we gather a possible answer? and how do we determine the best answer?

System description



Indexing

- Lucene
- Size of swedish wikipedia? Swedish vs english?
- Paragraphs vs Articles?
- Stemming or not?

Ranking

- Lucene ranks paragraphs with bm25
- Nouns are ranked from occurrences and bm25-score
- Limit to 100 nouns

$$\text{nounrank} = \sum(\text{bm25score} * \text{occurances})$$

Reranking

- Categorize question using libshorttext based on question set

```
location    Vad heter Sveriges huvudstad?
...
```

- Liblinear model based on previous results

T/F F1:Question words F2:Answer words F3:Question categories

```
1    1:1 7:1 166:1   5677:1  8001:0.6    8002:0.1        8003:0.1
…
```

Given a question and answer, predicts probability that this answer is correct.

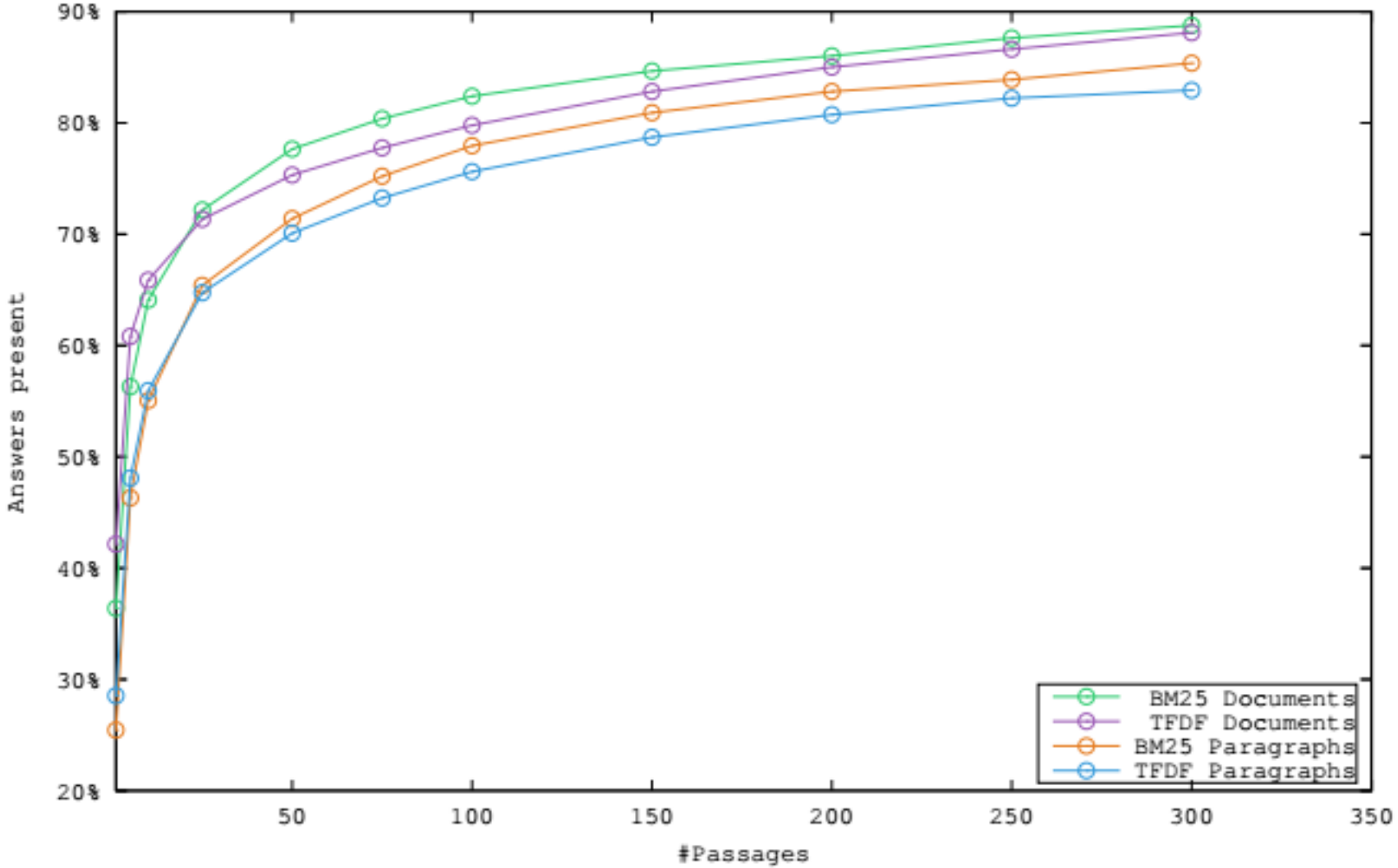
Results

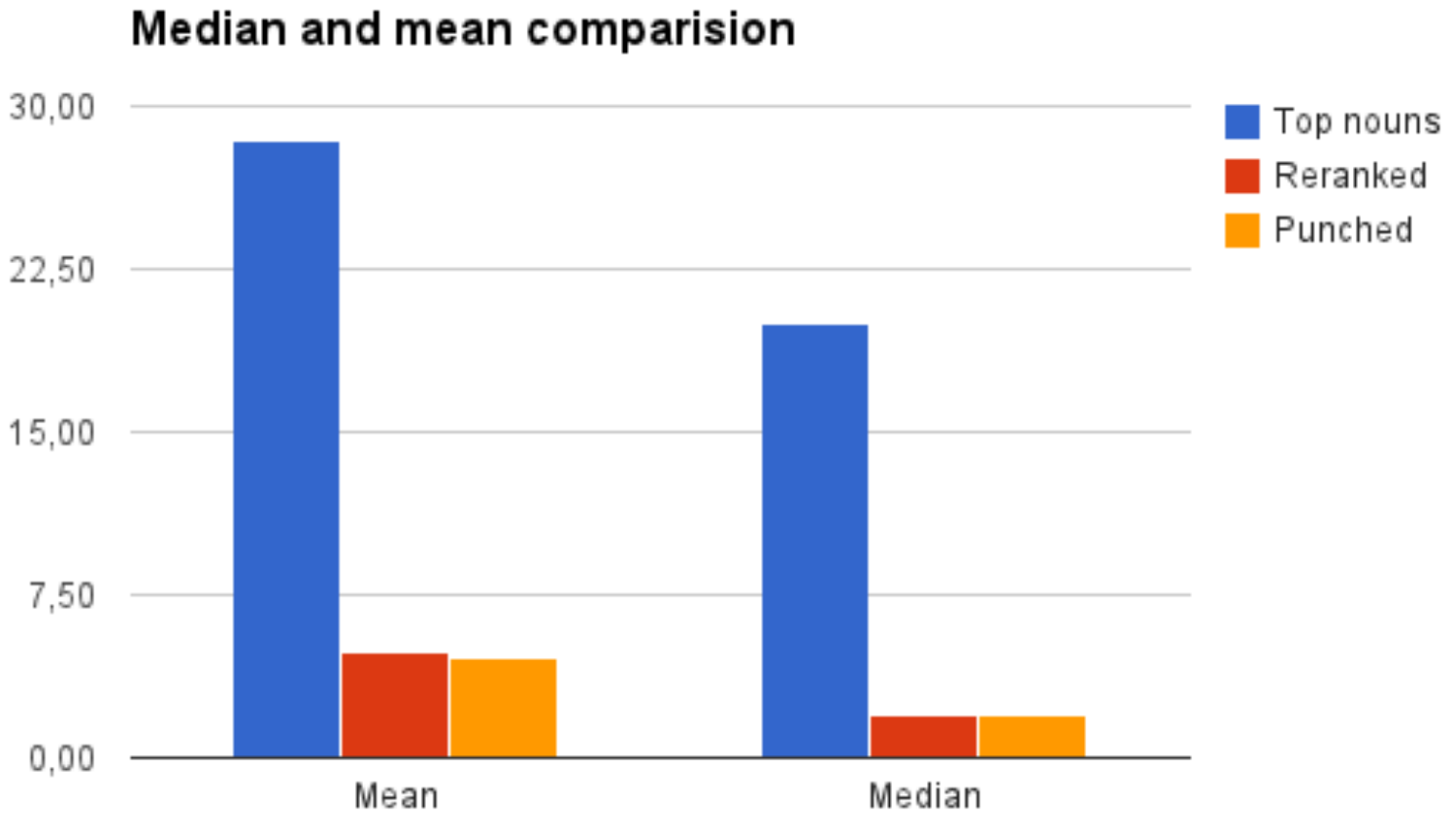
- Answer is present in passages
  #passages
  BM25 vs TFDF
  Paragraphs vs Documents

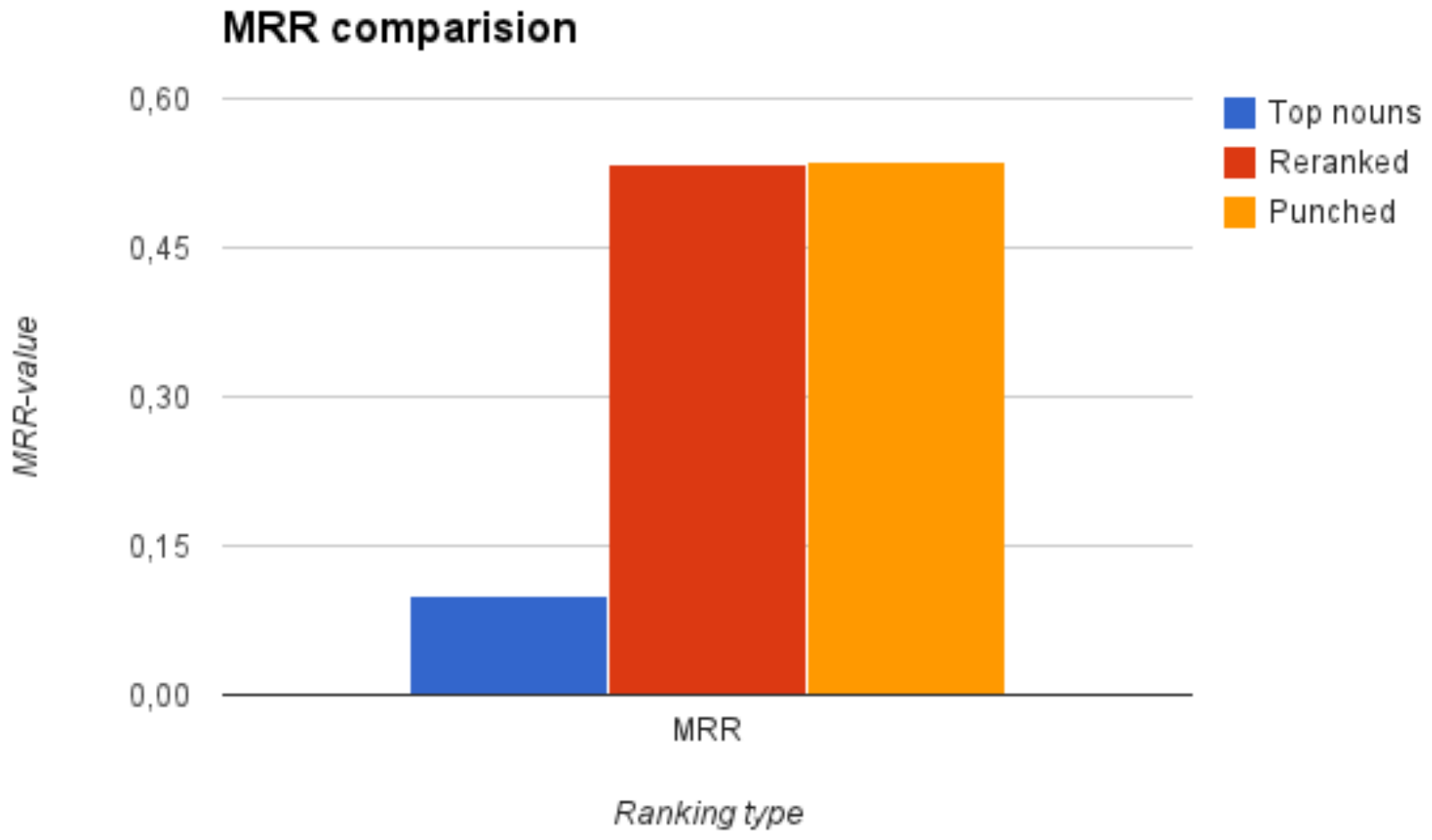
- Median rank among top nouns

- Median rank after rerank

- Median rank after punch

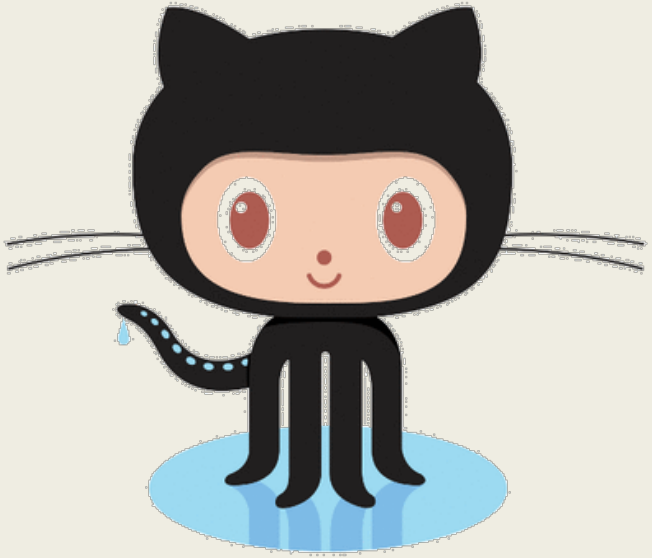

90%
80%
70%
60%
50%
40%
30%
20%
Answers present
50
100
150
200
250
300
350
#Passages
BM25 Documents
TFDF Documents
BM25 Paragraphs
TFDF Paragraphs


Median and mean comparision
30,00
22,50
15,00
7,50
0,00
Mean
Median
Top nouns
Reranked
Punched


MRR comparision
0,60
0,45
0,30
0,15
0,00
MRR-value
MRR
Ranking type
Top nouns
Reranked
Punched

Open source!

https://github.com/3amice/minerva



DEMO

Minerva