

Question Answer System Deliverable #2

Jonggun Park
Maria Antoniak


Haotian He
Ron Lockwood



System architecture

Two modules:

- ◊ Indexing
- ◊ Querying
 - ◊ query processing
 - ◊ passage retrieval
 - ◊ answer processing/ranking



Document Indexing/Retrieval

- ◊ Apache Lucene
- ◊ Two indices
 - ◊ Full text (used for idf calculations)
 - ◊ Paragraphs (used for scoring results)

Query processing

~~Why did Chuck Norris~~
uppercut a horse?



chuck, norris, uppercut, horse

Search

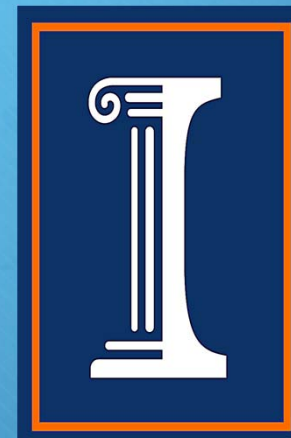
Query processing

+ POS

+ NER

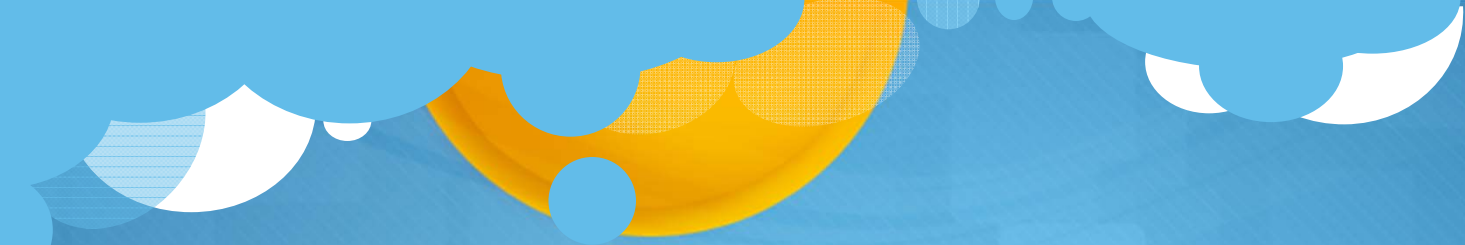
+ Chunking

+ Stemming





Chuck Norris uppercut a horse to
make a giraffe



Answer Extraction/Processing

- ◊ Initial solution is a redundancy-based strategy
- ◊ With one big difference
 - ◊ Instead of using web queries for snippets
 - ◊ We are using results (top 100) from a Lucene query





Answer Extraction Details

Input to the Extraction Engine

- Query word list
- Stop-word list
- Focus-word list (e.g. meters, liters, miles, etc.)
- Passage list – the paragraph results of the query

1. N-gram generation and occurrence counting
2. Filtering out stop words and query words
3. Combining unigram counts with n-gram counts
4. Weighting candidates with idf scores
5. Verifying candidates in documents

Lin, J. 2007. *An exploration of the principles underlying redundancy-based factoid question answering*. Penn Plaza, Suite 701, New York, NY.



D2 Results

◊ Strict = 0.01

◊ Lenient = 0.064

Low results... but improvements are coming!



Future work

- ◊ NER
- ◊ Web boosting
- ◊ Query/answer classification



Thank you!

Questions?