

# QA Baseline

Arjun Bhalla, Laurel Hart, Kathleen Kamali

# System Overview

- Query Processing
- Document Retrieval
- Answer Extraction

# Query Processing

- Remove stop words and punctuation
  - Stop word list
- Originally removed words shorter than 3 char
  - “UN”
- Each query considered separately (no series)

# Document Retrieval

- Used Lucene
- Built index storing doc path, number and text
- Searched over the text fields

# Answer Extraction

- Ngrams taken from each line of document
- Compared with query words
- Possible answers taken due to similarity of query
- Then ranked

# Results

Strict	$5.18134 \times 10^{-4}$
Lenient	$5.5568 \times 10^{-4}$

# To Be Improved

- Answer extraction
- Indexing paragraphs
- Training stop words list for query processing