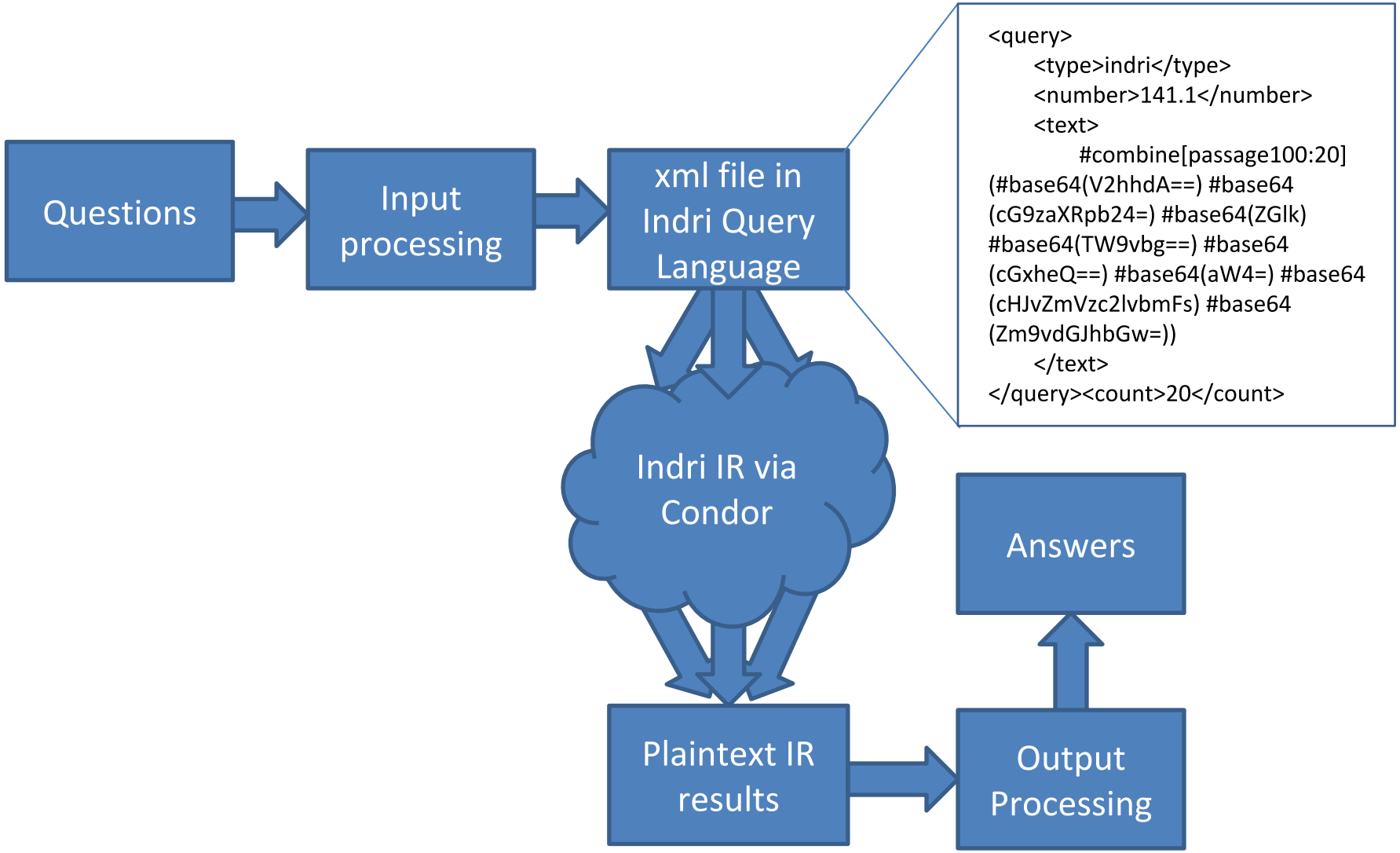
LING 573 D2
Indri Baseline Approach to QA

Melanie Bolla, Woodley Packard, and T.J. Trimble

System Architecture



System Architecture

Overview

- Process TREC input (with BeautifulSoup) and output an xml file in Indri format with question and target words → IndriRunQuery on the resulting xml file → output processing to select the 250 byte “answer”
- Indri processing distributed via Condor
- wrapper script for execution

System Architecture

Query Processing

- XML parsing with BeautifulSoup for question extraction
- remove question marks
- drop non-factoid questions
- Indri Query format:
  - #combine(): each term is weighted equally and combined in an “or” fashion
  - #base64: gets rid of punctuation problems

System Architecture

Retrieval:

- IndriRunQuery
- [passage100:20]; evaluates query on each sequence of 100 words, window slides 20 words at a time
- Takes ~30 seconds per query
- Several hundred queries for devtest
- 3 hours serial; 8 minutes distributed
- Top 20 results

System Architecture

- Answer candidate extraction
  First 250 byte window of IR results
- Reranking:
  - Tried n-gram based reranking technique
    - Didn’t help (yet)

Results

Results over DevTest

Query Formulation	Lenient Score	Strict Score
Unordered Question Words	0.1319	0.0753
Unordered Question Words + Target Words	0.1755	0.1073

Issues

- Passage selection
- Reranking
- Indri not retrieving passages as expected? – “printSnippets”.
  - Post-deadline improvements to Indri invocation mechanics:

    Lenient: 0.2204, Strict: 0.1395

Issues

- Stop words given too much weight
- Bag-of-words approach has limitations
- Redundancy-based answer reranking strategy not improving results
- Article headers

Error Analysis

2 Question Sets analyzed, 9 questions

Error Type	Percentage of Errors
Query Expansion	56%
Passage Selection	22%
Answer Processing	11%
Reranking	11%

Successes

- Base 64 conversion to help with punctuation
- Produces answers and supporting docs for all questions
- Including the target in the query as opposed to using the question alone (bag-of-words approach) improved our results

Influential Related Reading

- Class reading on Indri: http://sourceforge.net/p/lemur/wiki/Home/