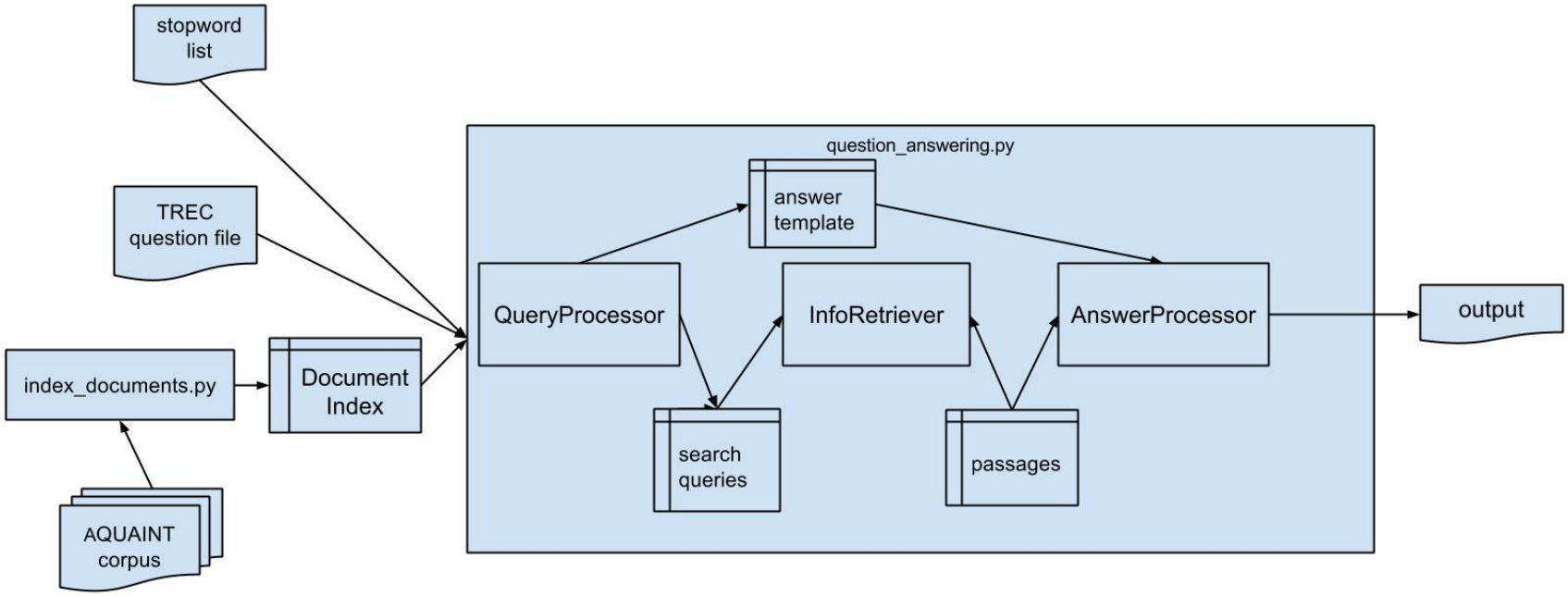
Baseline System Overview

Claire Jaja, Andrea Kahn and Clara Gordon

System Overview

- Use Indri/Lemur to create document index
  - Two indexes: Porter stemmer; Krovetz stemmer
- Use Beautiful Soup to extract questions from TREC XML document
- Generate Question objects that store question text, target, and ID
- Pass questions into pipeline one-by-one
- Print answers output by pipeline to output file

System Architecture



Query Processing

- Generate answer template
  - Currently, just includes a set of “words” (question/target tokenized on whitespace, with punctuation removed)
- Generate search queries, each of which consists of a set of weighted search terms
  - Currently, terms are just words in the question/target and weights are just counts
- Each search query itself has a weight
  - Currently, just one query per question; in future development, plan to generate multiple weighted queries
- Pass query to Passage Retrieval module
- Pass answer template to Answer Processing module

Passage Retrieval

- First pass: pymur
  - Only returned stemmed passage results
- Second pass: Indri/Lemur from command line
  - 100 snippet results passed on to answer processing
  - -printSnippets option
  - Other options: stopword list, combine query terms
  - Problems!
    - Only one snippet per document
    - Low quality and relevance
- Third pass (* in progress!)
  - Java module, accessing Indri API directly
    - (Thanks Woodley!)
  - Should allow us to access passage text directly

Answer Extraction and Ranking

- Extract answers
  - n-grams up to 4 from NLTK tokenized passages
  - score = sum of negated inverse of retrieval score of passages
- Filter answers
  - remove answers that stop or end with stopword
  - remove answers that contain any query words or punctuation
- Combine answers
  - update scores
  - new score = current score + sum of scores of unigrams contained
- Reweight answers
  - remove answers that don't occur in more than one passage
  - later will enact constraints on answer type
- Rank answers
  - rank by score, return top 20

Guess the Question!

142.4 D2 XIE20000928.0139 Conservatory of Music began
142.4 D2 XIE19990830.0092 Communist Party of China
142.4 D2 XIE20000928.0139 Central Conservatory of Music
142.4 D2 XIE19990916.0102 Republic of China.
142.4 D2 XIE19990811.0205 Republic of China
142.4 D2 XIE20000928.0139 Conservatory of Music
142.4 D2 XIE19990913.0201 founding of New China
142.4 D2 XIE19991001.0257 establishment of Sino-Russian diplomatic
142.4 D2 NYT20000113.0020 women 's professional sports
142.4 D2 NYT20000113.0020 women 's professional

Guess the Question!

142 target: “LPGA”

142.4 question: “When does the LPGA celebrate its 50th anniversary?”

142.4 D2 XIE20000928.0139 Conservatory of Music began

142.4 D2 XIE19990830.0092 Communist Party of China

142.4 D2 XIE20000928.0139 Central Conservatory of Music

142.4 D2 XIE19990916.0102 Republic of China.

142.4 D2 XIE19990811.0205 Republic of China

142.4 D2 XIE20000928.0139 Conservatory of Music

142.4 D2 XIE19990913.0201 founding of New China

142.4 D2 XIE19991001.0257 establishment of Sino-Russian diplomatic

142.4 D2 NYT20000113.0020 women 's professional sports

142.4 D2 NYT20000113.0020 women 's professional

Guess the Question!

192.2 D2 APW19980710.0900 independence of northern Spain
192.2 D2 APW19981128.0652 northern Spain and southern
192.2 D2 APW19981005.0486 northern Spain and southwestern
192.2 D2 APW19981005.0486 territory in northern Spain
192.2 D2 APW19981128.0652 parts of northern Spain
192.2 D2 APW19981128.0652 Spain and southern France
192.2 D2 APW19981028.0971 provinces of northern Spain
192.2 D2 APW19981028.0971 northern Spain and parts
192.2 D2 APW19981128.0652 Spain and southern
192.2 D2 APW20000717.0216 independent homeland in northern

Guess the Question!

192 target: “Basque ETA”

192.2 question: “Approximately how many people has ETA killed?”

192.2 D2 APW19980710.0900 independence of northern Spain

192.2 D2 APW19981128.0652 northern Spain and southern

192.2 D2 APW19981005.0486 northern Spain and southwestern

192.2 D2 APW19981005.0486 territory in northern Spain

192.2 D2 APW19981128.0652 parts of northern Spain

192.2 D2 APW19981128.0652 Spain and southern France

192.2 D2 APW19981028.0971 provinces of northern Spain

192.2 D2 APW19981028.0971 northern Spain and parts

192.2 D2 APW19981128.0652 Spain and southern

192.2 D2 APW20000717.0216 independent homeland in northern

Results

System	Strict	Lenient
Baseline (D2)	0.00511	0.02894

Lots of room for improvement...

Issues

- Poor quality of passages returned by Indri
- Duplicate documents cause less relevant documents to be too heavily weighted
- Query words not always matched in answer for filtering, due to stemming
- Tokenization issues with NLTK in passages (not all punctuation separated off)

Next Steps

- Query Processing
  - Generate multiple queries (e.g., query expansion using WordNet)
  - Question classification for better answer template generation
  - Removal of question words
  - Implement query term and overall query weighting
- Info Retrieval
  - Explore different methods of stemming
  - Explore using Indri API to return unstemmed passages, instead of stemmed passages or snippets (in progress!)
  - Used Indri Query Language with named entities
  - Attempt web boosting
- Answer Processing
  - Reweight answers based on question classification
  - Clever selection of doc ID as answer source

Thank you!