Milestone 2

Team 08

M1 Results

	doc1	doc2	doc3	doc4
PMI	0.22	0.77	0.33	0.2
Co-ref+PMI	0.4	0.5	0.4	0.2

- 11/40 give different choice: 4 Incorrect->Correct; 3 Correct->Incorrect; 4 remain incorrect.
- Our approach can find at least 1 candidate sentence for each question, so we answer all the questions. (3 questions are not answered in baseline system, because it doesn't find any candidate sentences.)

M1 Error Analysis --> M2 Changes

- Incorrect->Correct:
  - Enhanced search leads to more relevant documents. ‘null’ candidate sentences too are populated (e.g. D3Q3).

- Correct->Incorrect:
  - More noise is added due to query expansion and potentially incorrect coreference. Correct answers lose out only slightly in most of the cases.
  - **M2 Enhancement**: Updated Solr query (was: NE + NP + coref + synonyms) and added raw text (now: NE + NP + coref + synonyms + text); modified Solr weights to give higher importance to NEs and NPs.

M1 Error Analysis --> M2 Changes

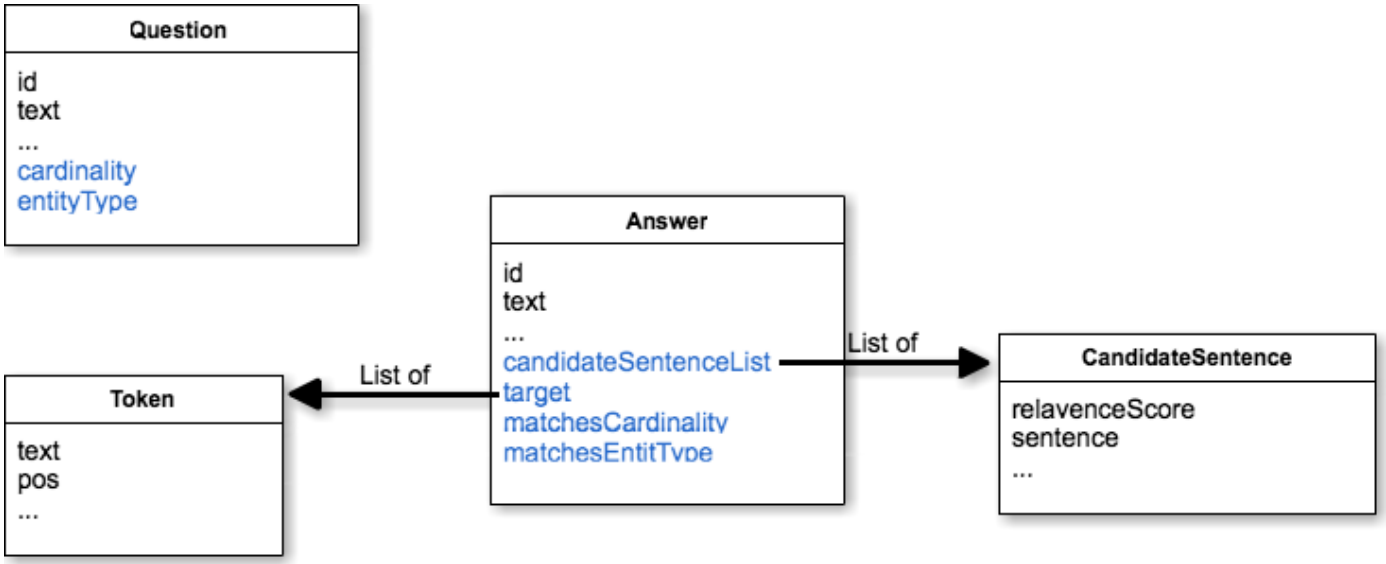
- Incorrect->Incorrect:
  - Need an algorithm to eliminate obviously wrong answers for certain question types (“How many...”, “When...”). NP-based PMI (for answer scoring) does not cover answers which are adjective-based (e.g. “longer”, “shorter”).
  - **M2 Enhancement**: Added flags to check for the Question cardinality and Answer cardinality, to see if they match. Added flags to check for Question target entity type (integer, double, etc) and Answer entity type, to see if they match. Used these flags as components within a combined answer scoring function.

  - Often, words from the question and words from the answer can be found within the same sentence.
  - **M2 Enhancement**: Combined Question terms and candidate Answer terms into Question+Answer queries. Used the resulting candidate sentences to compute PMI scores which were incorporated into a combined scoring function.

M1 Error Analysis --> M2 Changes

- And finally...
  - **M2 Enhancement**: Created a combined scoring function that incorporates multiple sources of information to give an aggregate score:
    - Distance Metrics
      - Question Cand. Sentences vs. Answer → PMI (NP, NE, coref)
      - Question Cand. Sentences vs. Answer → Overlap (NP, NE)
      - Question Cand. Sentences vs. Answer Cand. Sentences → PMI (NP, NE, coref)
    - Flags
      - Answer.matchesQuestionCardinality → bonus / penalty for match / mismatch
      - Answer.matchesQuestionEntityType → bonus / penalty for match / mismatch

Type System Modifications



M2 Results

Incremental boost from enhancements:

	doc1	doc2	doc3	doc4	mean
M1	0.4	0.5	0.4	0.2	.375
M2 - v1	0.4	0.6	0.4	0.3	.425
M2 - v2	0.4	0.4	0.1	0.6	.375

Example Changes:

Question: In which cell line was the gene expression regulation of IDE characterized?(D2Q3)

Correct Answer: BV-2

M1 result: beta-amyloid

M2 result: BV-2

In M1, the candidate sentences that support “BV-2” rank lower because the synonym expansion has higher weight which gives higher rank to the sentences which support “beta-amyloid”

In M2, the candidate sentences that support “BV-2” rank higher than M1 because we reduce the weight for synonym expansion in our query

Next Steps: Milestone 3

Parameter Tuning

- Deep dive analysis to determine:
  - Optimal weights for Solr query.
  - Optimal weights for combining answer scoring components.
  - Optimal K-value for candidate sentence selection.
  - Optimal parameters in Noise Filter

Answer Scoring / Answer Selection:

- Based on analysis above, incorporate additional distance scores if needed.
- (Time permitting): Build function to handle “None of the Above” cases.