TEAM 10

Milestone 2 Report 11-791 Project

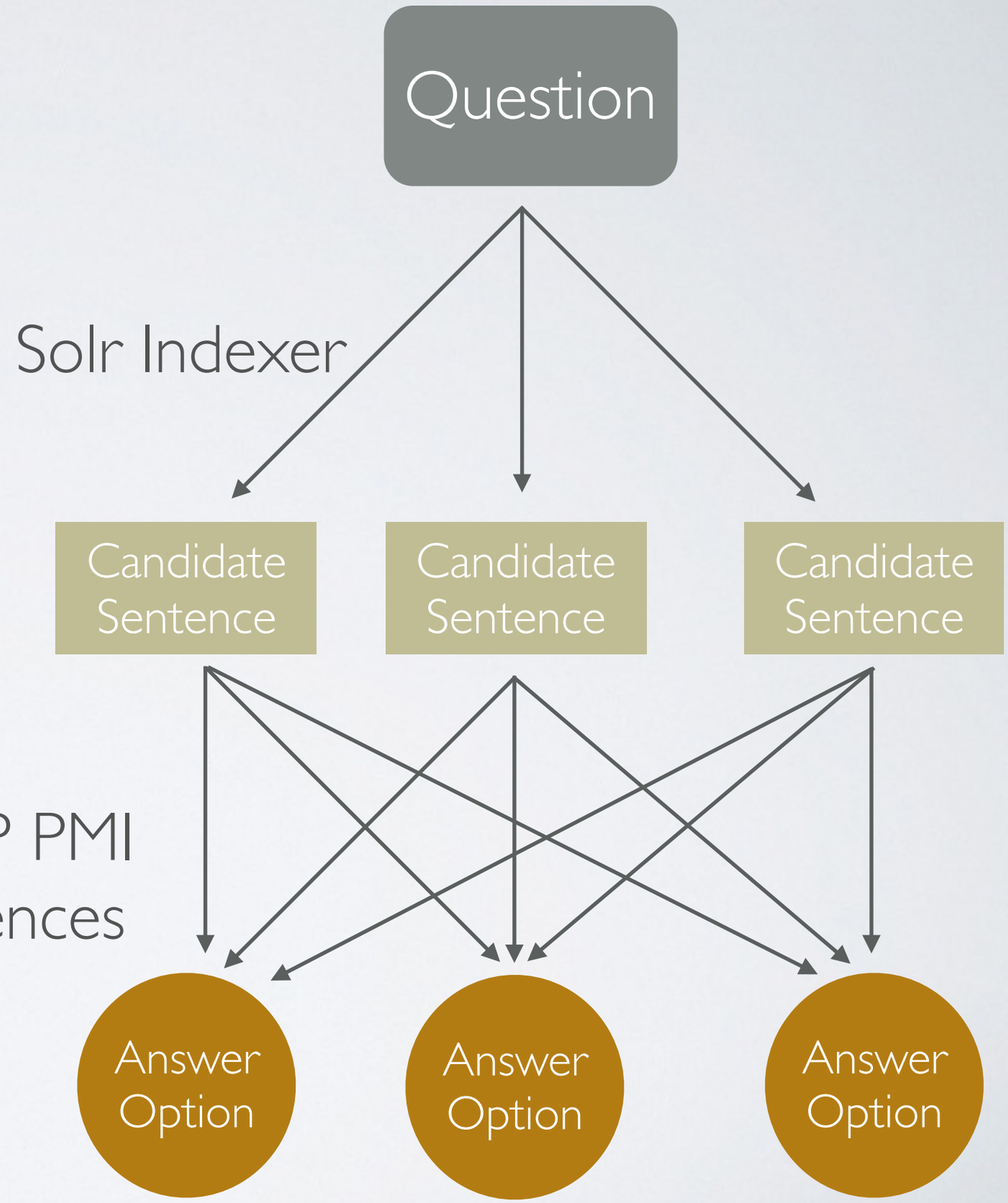
Error Analysis

Nov. 13, 2013

Troy Hua Chenyan Xiong Vinay Vemuri Bo Ma

https://github.com/troyhua/hw5-team10

BASELINE FRAMEWORK

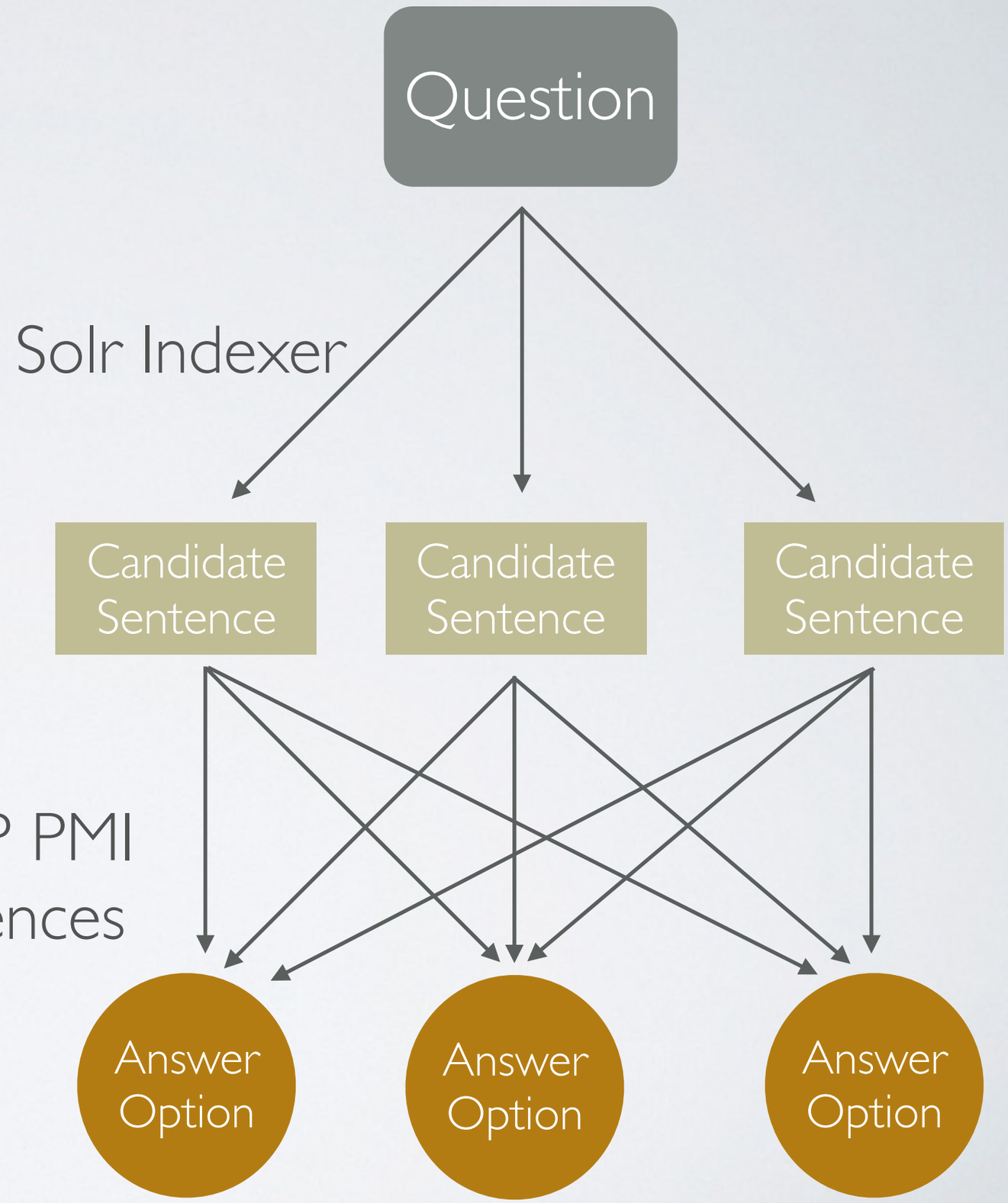


Joint Voting
NP Similarity and NP PMI
with candidate sentences

BASELINE FRAMEWORK

Candidate Sentences may not related to the answer

Joint Voting
NP Similarity and NP PMI
with candidate sentences

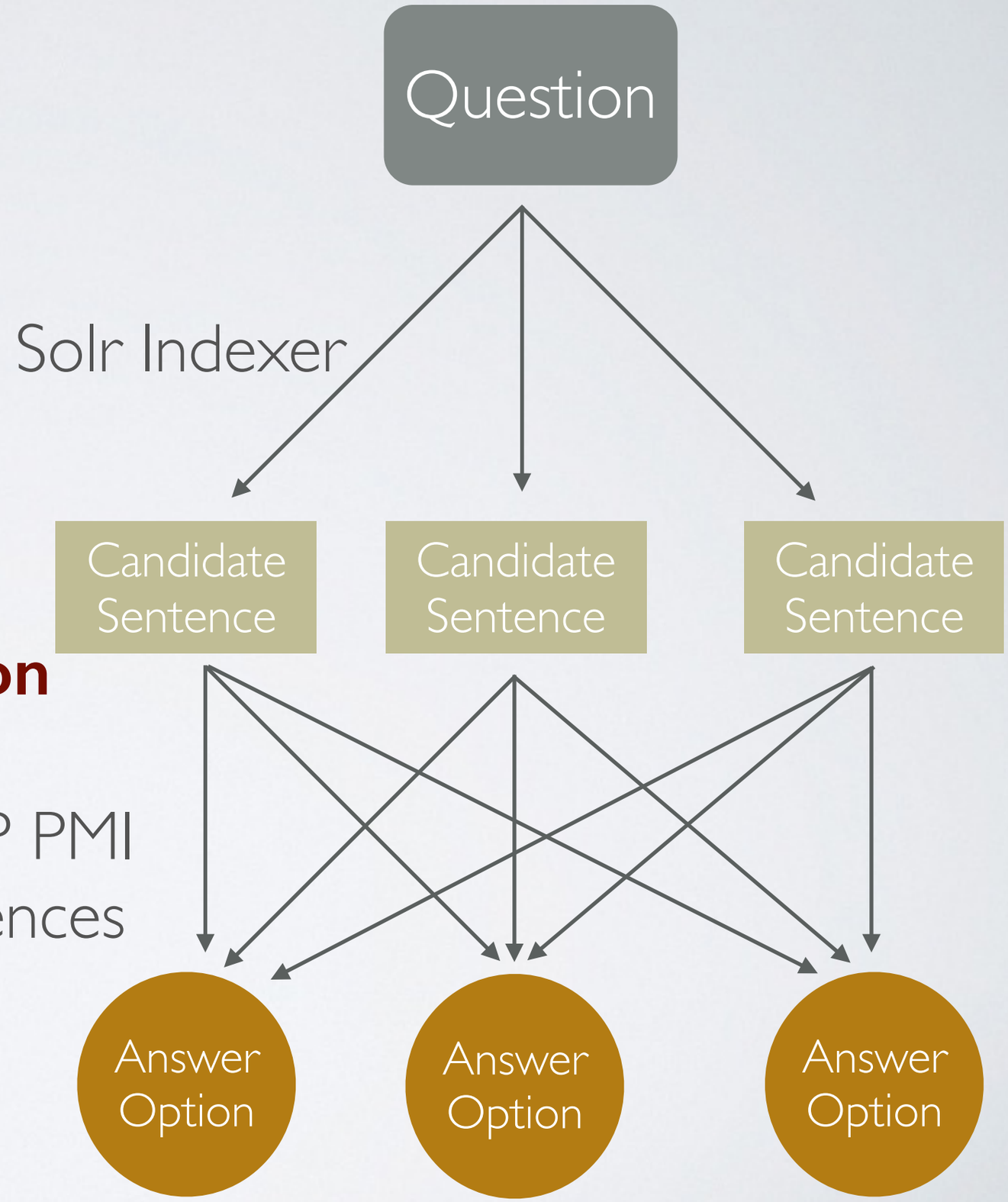


BASELINE FRAMEWORK

Candidate Sentences may not related to the answer

Answer Option may be relevant to the candidate sentence, but not the question

Joint Voting
NP Similarity and NP PMI
with candidate sentences



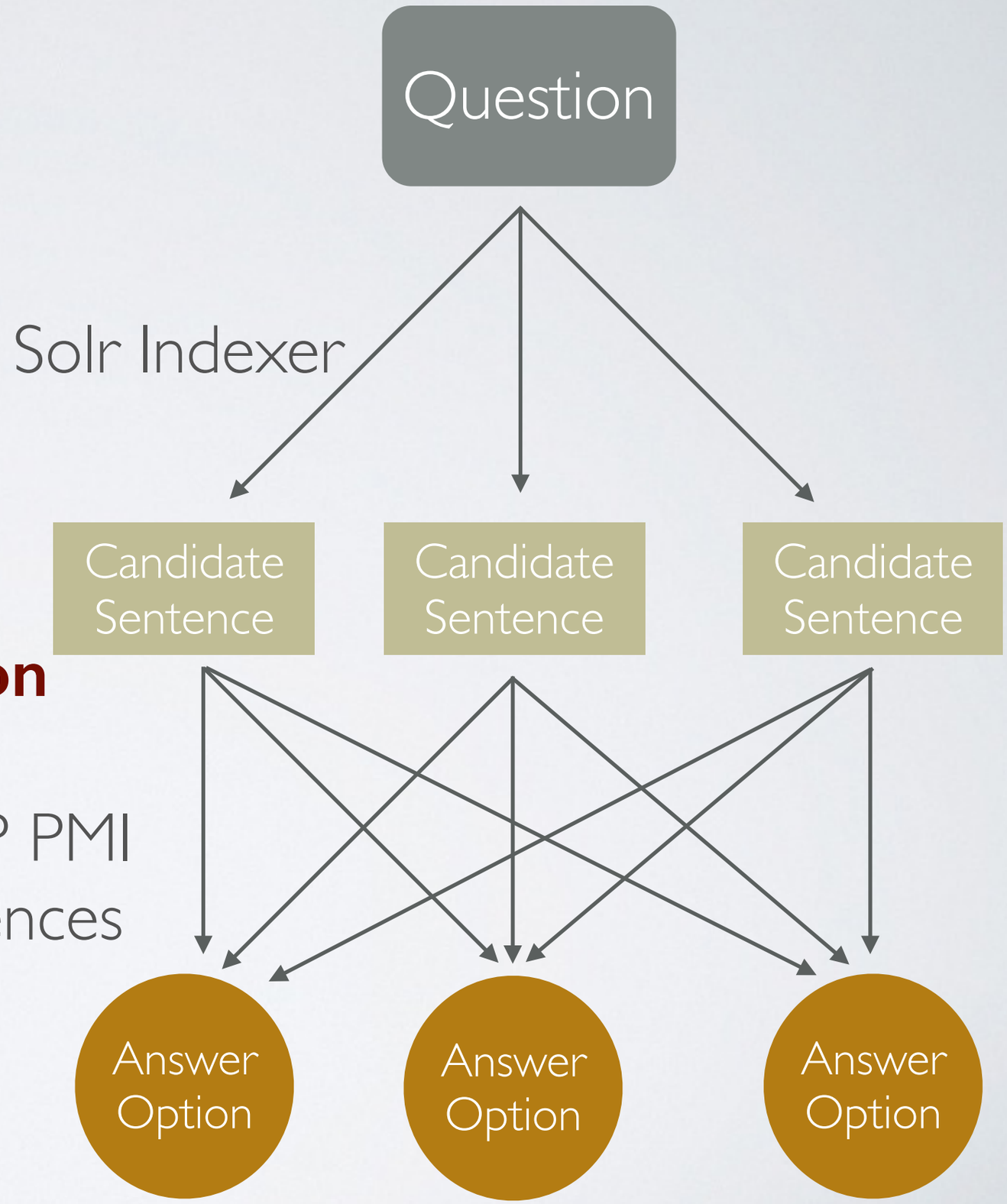
BASELINE FRAMEWORK

Candidate Sentences may not related to the answer

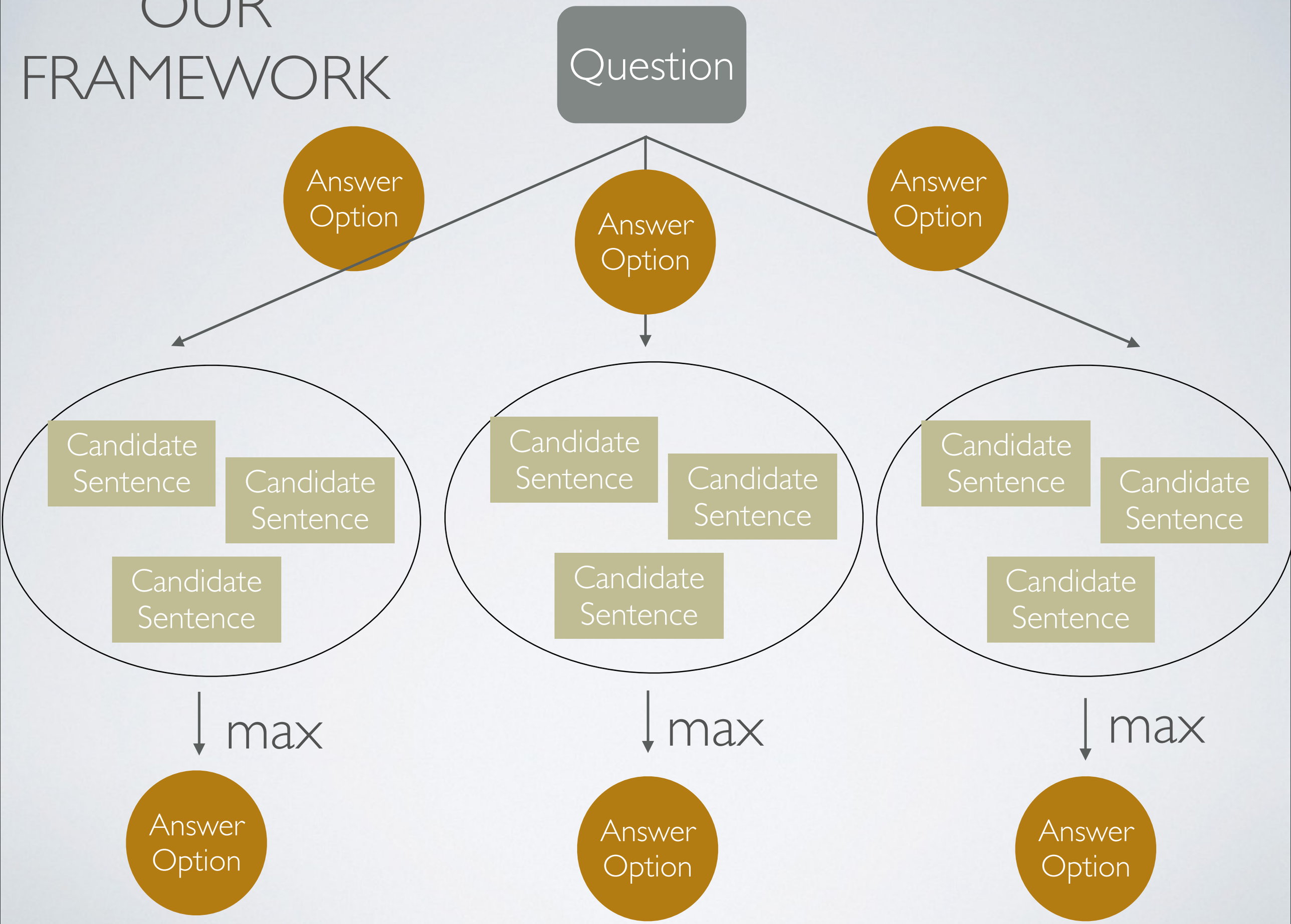
Answer Option may be relevant to the candidate sentence, but not the question

Joint Voting
NP Similarity and NP PMI
with candidate sentences

Joint Voting tends to select more frequent terms



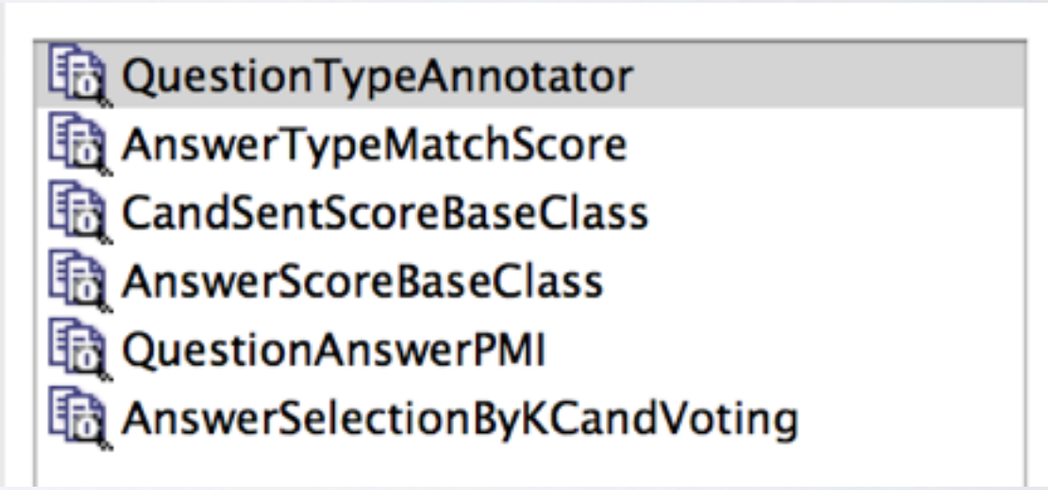
OUR FRAMEWORK



PERFORMANCE

- Performance on **2012** test set:
  - **Baseline**, without PMI
    - **20%** (3/10, 1/10, 1/10, 3/10)
  - **M1** Results, change the framework
    - **27.5%** (3/10, 4/10, 2/10, 2/10)
  - **M2** Results, details in the following slides
    - **47.5%** (3/10, 6/10, 7/10, 2/10) - It must be highly overfitting…

Current Pipeline



DETAILS

- Change Candidate Sentence selection from **Solr Indexer** to our own similarity metrics combining **NP, NER, Unigram** - **cosine similarity** and **match counts**

- Answer score = (**answer-candidate sentence** relevance) X (**candidate sentence-question** relevance)

- **Question-answer** PMI score instead of **candidate sentence - answer** PMI score

- **Question Type** match score, QUANTITY, REASON, CHOICE, TYPE, etc..

CANDIDATE SENTENCE SELECTION

- **Noun Phrase** and **Named Entity** are much much better features compared with **Unigram**

- The recall of the annotation for **NE** is not that good. We have different weights and methods for different features.

  - Cosine for similarity between Unigram of <question, sentence>

  - Match count for NER between <answer, sentence>

  - Dot Product for NP between <answer, sentence>, <question, sentence> with normalization only over the sentence NP list.

- Hard to do this in Solr query framework.

ANSWER SCORING

- Answer score = (**answer-sentence** relevance) X (**sentence-question** relevance)

- Similar similarity score function with the previous steps

QUESTION-ANSWER PMI

- Local features: Instead of co-occurrence of answer option and candidate sentence, we use the co-occurrence of answer-question pair.

- Combine NER and NP PMI

ERROR ANALYSIS

FOR THE NEXT STEP

SEMANTIC MATCHING

ERRORS

- **Question**: Which ***technique*** was used to determine the cellular locations of the CLU1 and CLU2 gene products?
  - **Wrong** Answer: intracellular and secreted *(not a technique, even a noun)*
  - **Correct** Answer: immunofluorescence experiments *(is a technique, at least a noun)*
- **Question**: Which ***hormone*** can control the expression of CLU isoforms?
  - **Wrong** Answer: real-time PCR *(not a kind of hormone)*
  - **Wrong** Answer: cDNA *(not a kind of hormone)*
  - **Correct** Answer: androgen *(a kind of hormone)*

SEMANTIC MATCHING

SOLUTION

- Using semantics/word correlation to determine the target relationship between entities
  - Should consider both semantic correlation and grammatical consistent (POS, number).
- For normal terms, word-net.
- For Bio terms, **PMI** and **distance**
  - **e.g.** "…**Fibrous Proteins** form muscle fiber, tendons…for example, **Actin**, **Arp2/3**, **Collagen**…."

SIGNIFICANT **TERM** DETECTION

ERROR

- Question: What **compartments** inside the **cell** contain **cluster** in **proteins**?

- Bad similarity scores for both candidate sentence and answer options

- **"Cell"** and **"Proteins"** dominate the NP/NER scores

SIGNIFICANT **TERM** DETECTION

SOLUTION

- Detect by **inverse-sentence-frequency**.

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{t \in d, d \in D\}|}$$

- Cell, proteins are very frequent in document, but cluster is not

- Use idf as weight for different terms in the question PMI

SIGNIFICANT **TERM** DETECTION

SOLUTION

Hard rules for different question type,

- Question: How many **residues** does the CLU2 protein sequence have?

- Rule: The first NP after "how many" must appear in the candidate sentence.

DOCUMENT CLEANING

Question: How many residues does the CLU2 protein sequence have?

Wrong Answer: 6

Best Candidate Sentence:
Although the primary role of clusterin in AD is unclear, **CLU** is implicated in AD by several lines of evidence, including (I) **CLU** mRNA and clusterin **protein** is increased in AD [5], [**6**], (ii) clusterin is a component of plaques [4], [5], [7], (iii) clusterin modulates AD-related pathways such as inflammation and…

SENTENCE WEIGHT FROM LOCATION IN THE DOCUMENT

- Candidate sentences should not be in the reference part.
- First sentence and last sentence of a paragraph are more important.

SUMMARY

- Performance Improved to 47.5%
  - Change Solr Indexer to more customized similarity metrics
  - Consider answer when selecting candidate sentences
  - Consider question when ranking answers
  - Question-Answer PMI score
  - Question Type match score
- Next steps:
  - Semantic Correlation
  - Significant terms detection
  - Location Weight for candidate sentence
  - Document Cleaning

TEAM 10

Milestone 2 Report 11-791 Project

Error Analysis

Nov. 13, 2013

Troy Hua Chenyan Xiong Vinay Vemuri Bo Ma

https://github.com/troyhua/hw5-team10