Milestone 1

Team 9: Napat Luevisadpaibul, Dishan Gupta, Callie Vaughn, Wenyi Wang

Baseline

The baseline system has an average accuracy of 17.5% (c@1 = 0.1925) across the four input files.

Baseline Accuracy on Input Files					
	XML 1	XML 2	XML 3	XML 4	Mean
% correct c at 1	2/10 = 20% 0.22	1/10 = 10% 0.11	1/10 = 10% 0.11	3/10 = 30% 0.33	7/40 = 17.5% 0.1925

Our Improvements

- Remove systematically invalid answers
  - Questions beginning with “How many” must have numeric answers

ex.

```
<q_str>How many mutations relevant for familial forms of Alzheimer's disease have been detected for the PSE
<answer a_id='1' >13</answer>
<answer a_id='2' >42</answer>
<answer a_id='3' discard="Yes">P436Q</answer>
<answer a_id='4'  correct="Yes">185</answer>
<answer a_id='5' discard="Yes">PSEN2</answer>
```

Our Improvements

- Remove systematically invalid answers
  - Questions beginning with “How many” must have numeric answers (This question probably wants an integer)

```
<q_str>How many residues does the CLU2 protein sequence have?</q_str>
<answer a_id='1'  correct="Yes">449</answer>
<answer a_id='2' discard="Yes">protein</answer>
<answer a_id='3' discard="Yes">82.3</answer>
<answer a_id='4' discard="Yes">52.5</answer>
<answer a_id='5' >6</answer>
```

Our Improvements

- Remove systematically invalid answers
  - Questions beginning with “How many” must have numeric answers (need some semantic analyzing to discard 60% and 70 year old)

ex

```
<q_str>How many persons worldwide are estimated to have a medical condition related to
<answer a_id='1' discard="Yes">LPR2</answer>
<answer a_id='2' discard="Yes">mitochondria</answer>
<answer a_id='3' discard="Yes">60%</answer>
<answer a_id='4'  correct="Yes">more than 10 million</answer>
<answer a_id='5' discard="Yes">70 years old</answer>
```

Our Improvements

- Remove systematically invalid answers
  - Questions beginning with “What are” must have plural answer.

ex.

```
<q_str>What are the sst receptors that are expressed on rat astrocytes?</q_str
<answer a_id='1' >SSTR-2, SSTR-3 and SSTR-4</answer>
<answer a_id='2'  correct="Yes">SSTR-1, SSTR-2 and SSTR-4</answer>
<answer a_id='3' discard="Yes">somatostatin</answer>
<answer a_id='4' discard="Yes">microglia</answer>
<answer a_id='5' discard="Yes">rat</answer>
```

Our Improvements

- Remove systematically invalid answers
  - Questions beginning with “What is” must have singular answer.

ex.

```
<q_str>What is the major protease produced by microglia responsible for degrading A?</q_s
<answer a_id='1' >Ab</answer>
<answer a_id='2' discard="Yes">cells</answer>
<answer a_id='3' >IDE</answer>
<answer a_id='4'  correct="Yes">extracellular</answer>
<answer a_id='5' >beta-amyloid</answer>
```

Our Improvements

- Remove systematically invalid answers
  - Questions beginning with “What” or “Which” followed by a noun phrase must have singular answers if the NP is singular and plural answers if the NP is plural

```
<q_str>What compartments inside the cell contain clusterin proteins?</q_str>
<answer a_id='1'  correct="Yes">ER and the Golgi apparatus</answer>
<answer a_id='2' discard="Yes">epitope tag</answer>
<answer a_id='3' discard="Yes">anibody</answer>
<answer a_id='4' discard="Yes">secretory pathway</answer>
<answer a_id='5' discard="Yes">secreted</answer>
```

Our Improvements

- Implement PMI scoring
  - What is PMI?
    - PMI = p(x,y) * log(p(x,y)/(p(x)*p(y)))
  - Why PMI?
    - It is a measure of the relative co-occurrences of two quantities. We applied it to named-entities in candidate sentence and answer choice.

Results

Accuracy on input files with improvements to baseline						
		XML 1	XML 2	XML 3	XML 4	Mean
Baseline	% correct c at 1	2/10 = 20% 0.22	1/10 = 10% 0.11	1/10 = 10% 0.11	3/10 = 30% 0.33	7/40 = 17.5% 0.1925
PMI Scoring	% correct c at 1	4/10 = 40% 0.44	2/10 = 20% 0.22	4/10 = 40% 0.44	1/10 = 10% 0.11	11/40 = 27.5% 0.3025
PMI + Removed Answers	% correct c at 1	5/10 = 50% 0.55	2/10 = 20% 0.22	4/10 = 40% 0.44	1/10 = 10% 0.11	12/40 = 30% 0.33

Future Work

- Use synonyms of name entities to find candidate sentences and use synonyms in candidate sentences to score answer choice. Our synonym system can support synonyms of phrases too.

  ex. Red blood cells -> erythrocytes

Future Work

- Implement system to discard unlikely answers automatically instead of doing it by hand