**Problem**

- Even largest available knowledge bases are incomplete.
- Relation extraction from unstructured data is one way to narrow the gap.
- Question-Answer pairs (QnA) are attractive data source for relation extraction, they contain information users are interested in.
- Existing approaches are typically based on various syntactic patterns and operate over individual sentences.
- However, often an answer is hard to understand without knowing the question.

**Approach**

(Question, Answer)

**Models**

1. **Baseline sentence-based model**
   - **Q:** Where was Isaac Newton born?
   - **A:** Born in Woolsthorpe-by-Colsterworth, England
   - **Features:** dependency tree and surface patterns
     - where <PER> born
     - where+born  // question word and main verb

2. **Sentence-based model with question features**
   - **Q:** Where was Isaac Newton born?
   - **A:** Isaac Newton (Woolsthorpe-by-Colsterworth)
   - **Features:** conjunctions of question and answer patterns
     - where <PER> born
     - where+advmod(born)
     - where+born // question word and main verb

3. **Question-Answer based model**
   - **Q:** Where was Isaac Newton born?
   - **A:** Woolsthorpe-by-Colsterworth
   - **Features:** conjunctions of question and answer patterns
     - where <PER> born A:LOC
     - Q{where}+advmod(born)nsubj+ <PER> A:LOC
     - Q: where + <PER> + born A:LOC

**Experiments**

- Yahoo! Answers dataset
  - 3.8M QnA pairs
  - 3.3M question clusters with an answer (19.6M QnA pairs)
- WikiAnswers dataset
  - 3.8M question clusters with an answer (19.6M QnA pairs)

**Results**

- We can achieve higher precision and recall by combining sentence-based and QnA-based models.
- QnA-based model extracts from 20-40% of triples not extracted by the sentence-based model.

**Error analysis (false positives):**
- ~40% due to entity linking problems
- ~16% cases require deeper understanding of the answer text
- ~8% cases contradict Freebase data
- ~33% are correct extractions and are missing in Freebase

**Conclusion**

- We proposed a model for relation extraction from QnA data that models the discourse of the pairs and can extract relations between entity pairs mentioned in question and answer sentences.
- Conducted experiments on 2 publicly available datasets show that the model can be effectively combined with existing sentence-based techniques and produces from 20-40% new relation triples.

This presentation was supported by the Google Faculty Research Award and NSF.