Relation Extraction from Community Generated Question-Answer Pairs

Denis Savenkov
Emory University
dsavenk@emory.edu

Wei-Lwun Lu
Google
weilwunlu@google.com

Jeff Dalton
Google
jeffdalton@google.com

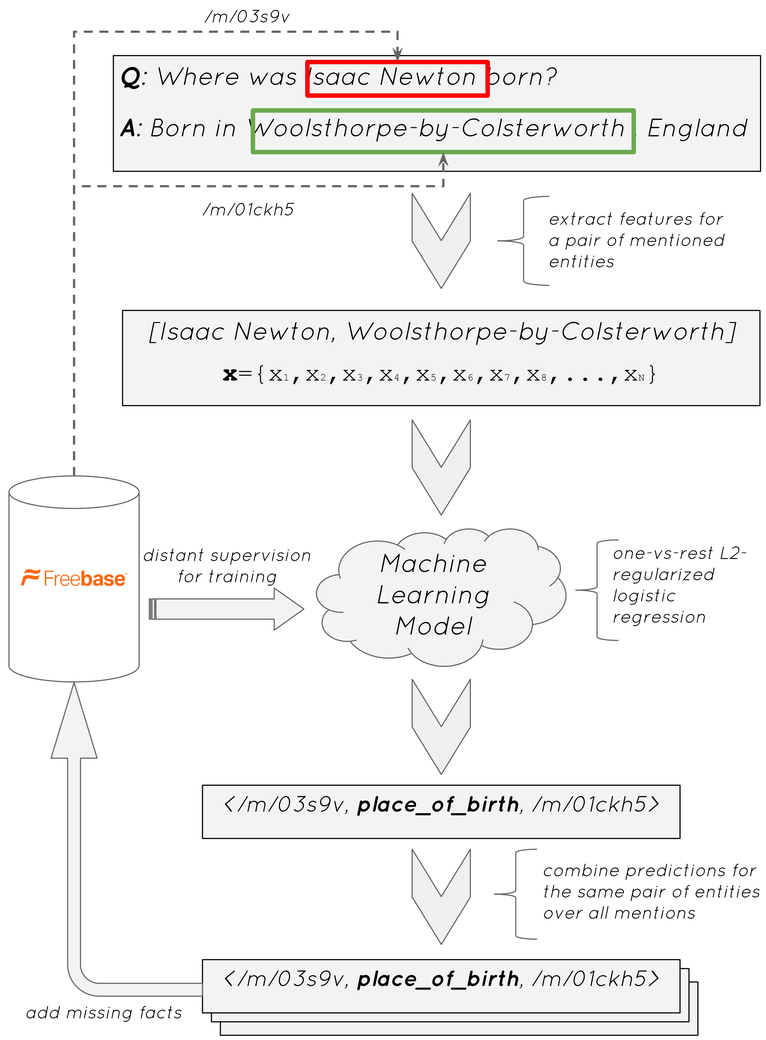
Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

Intelligent Information Access Lab (IRLab)
http://ir.mathcs.emory.edu

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



Models

1. Baseline sentence-based model

Q: Where was Isaac Newton born?
A: Isaac Newton was born in Woolsthorpe-by-Colsterworth

Features: dependency tree and surface patterns

- [+context] <PER>-nsubjpass→(born)←nmod-<LOC>[+context]
- [+context] <PER> be/VBD born/VBN <LOC> [+context]

2. Sentence-based model with question features

Q: Where was Isaac Newton born?
A: Isaac Newton (Woolsthorpe-by-Colsterworth)

Features: above + question patterns features

- 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

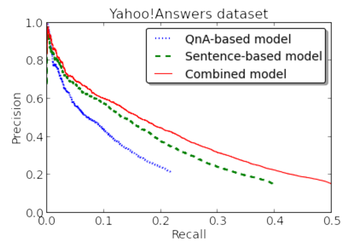
- Q: where <PER> born A:<LOC>
- Q:(where)→advmod(born)nsubj←<PER> A: <LOC>
- Q: where + <PER> + born A:<LOC>

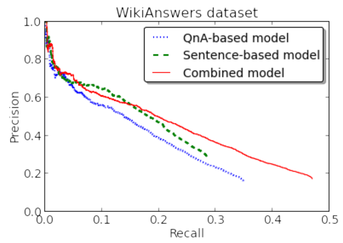
Experiments

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

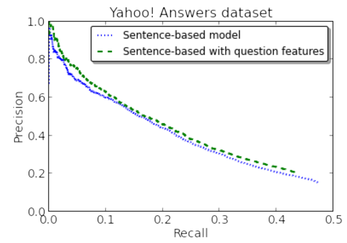
- NLP pipeline (dependency parsing, ner, coreference)
- entity linking (entity names and anchor text dictionary)
- split documents and relation triples for training/testing
- train a model using distant supervision
- evaluate the model by precision/recall on held-out triples

Results





We can achieve higher precision and recall by combining sentence-based and QnA-based models



Question features didn't help sentence-based model

	Yahoo! Answers dataset			WikiAnswers dataset		
	QnA	sent	comb	QnA	sent	comb
Number of correct extractions	3229	5900	7428	2804	2288	3779
Correct triples not extracted by other model	20.5%	56.5%	-	39.4%	25.8%	-

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.