Bootstrapping for Numerical Open IE

Swarnadeep Saha
Department of CSE, I.I.T. Delhi

Harinder Pal
Microsoft Corporation

Mausam
Department of CSE, I.I.T. Delhi

Introduction

Open IE

- An **Open Information Extraction(Open IE)** system extracts relational tuples from text:
  - without requiring a pre-specified relational vocabulary.
  - by identifying relational phrases and arguments from the sentences themselves.
- Early works like **ReVerb** (Etzioni et al., IJCAI 2011) extract verb-mediated relations.
- Subsequent works like **OLLIE** (Mausam et al., EMNLP-CoNLL 2012) have focused on increasing recall using bootstrapping over **ReVerb** extractions.
- **Open IE 4.2** (https://github.com/knowitall/openie) , a state-of-the-art open information extraction system is oblivious to the presence of numbers in arguments; thus misses important extractions and may not output the best numerical facts.

Closed Numerical IE

- Explicit numerical relation extractors like **NumberRule** (Madaan et al., AAAI 2016) extract relations where one of the arguments is a quantity.
- However, all are ontology-specific and do not directly apply to Open IE.

Our Contribution : Open Numerical IE

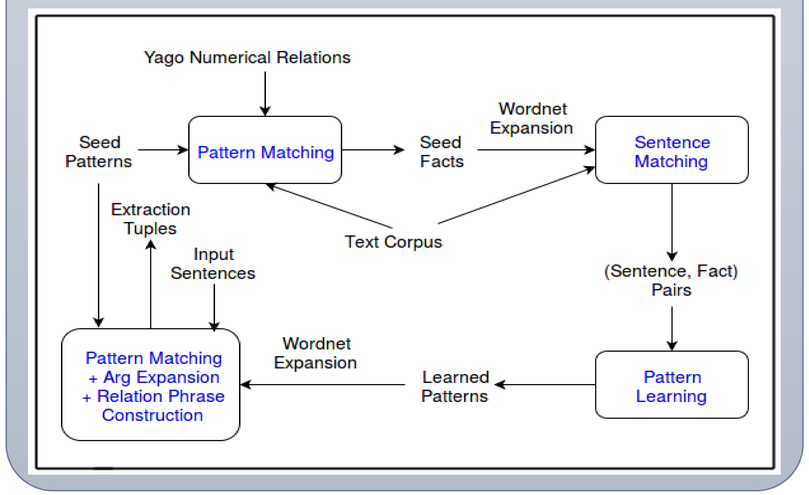
- We release the first system for open numerical extraction named **BONIE**.
- **BONIE** follows **OLLIE**'s design at a high level.
- Our customizations specific to Numerical IE:
  - Manually define a set of high-precision seed dependency patterns.
  - Develop heuristics to identify an informative bootstrapping set.
  - BONIE identifies implicit numerical relations from sentences.
- **BONIE** is freely available. **https://github.com/Open-NRE**

Seed Patterns

Seed Dependency Patterns

1. <(#is|are|was|were|been|be#verb)<(nsubj#{rel}#nnp|nn)<(prep#of|for#in)<(pobj#{arg}#nnp|nn|prp)>>(attr#{quantity}#.+)>>
2. <(#has|have|had|having#verb)<(dobj#{rel}#nnp|nn)<(prep##in)<(pobj#{quantity}#.+)>>(nsubj#{arg}#nnp|nn|prp)>>
3. <(#is|are|was|were|been|be#verb)<(nsubj#{arg}#nnp|nn|prp)(acomp|advmod#{rel}#jj|rb)<(npadvmod#{quantity}#.+)>>>
4. <(#has|have|had|having#verb)<(nsubj#arg#nnp|nn|prp)(dobj#quantity#.+)<(prep#of#in)<(pobj#{rel}#nnp|nn)>>>>
5. <(##verb)<(attr|acomp#{quantity}#.+)(nsubj#{rel}#nnp|nn)<(poss#{arg}#nnp|nn)>>>
6. <(#{rel}#verb)<(auxpass#is|are|was|were|been|be#verb)(nsubjpass#{arg}#nnp|nn|prp)(prep##in)<(pobj#{quantity}#.+)>>

BONIE Flow Diagram



Comparison of Open IE 4.2 and BONIE on some sentences

Sentence	Open IE 4.2	BONIE
Hong Kong's labour force is 3.5 million.	(Hong Kong's labour force; is; 3.5 million)	(Hong Kong; has labour force of; 3.5 million)
Microsoft has 100,000 employees.	(Microsoft; has; 100,000 employees)	(Microsoft; has number of employees; 100,000)
James Valley is nearly 600 miles long.	(James Valley; is; nearly 600 miles long)	(James Valley; has length of; nearly 600 miles)
Donald Trump is 70 years old.	(Donald Trump; is; 70 years old)	(Donald Trump; has age of; 70 years)
James Valley has 5 sq kms of fruit orchards.	(James Valley; has; 5 sq kms of fruit orchards)	(James Valley; has area of fruit orchards; 5 sq kms)

Generation of Seed Facts

- 6 manually written high-precision seed dependency patterns.
- Each dependency pattern encodes the minimal sub-tree of the dependency parse connecting the relation, quantity and argument in that sentence.
- {rel}, {arg} and {quantity} are placeholders for relation, argument and quantity headwords respectively.
- **BONIE** matches dependency parse of a sentence with a pattern to generate seed facts.
- A seed fact is of the form **(arg headword; relation headword; quantity; unit).**
  e.g. (India; population; 1.2 billion; null)
- To remove generic and noisy seed facts, **BONIE** keeps only those seed facts which are common with numerical facts in **Yago KB**.
- **BONIE** uses **Wordnet** expansion to convert non-nominal relations into nominal ones.
  e.g. (Brown; tall; 13; inches) -> (Brown; height; 13; inches)

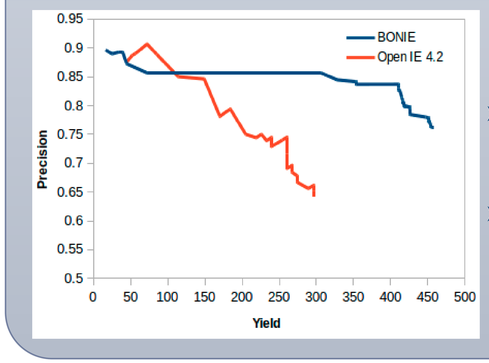
Bootstrapping

- **BONIE** finds sentences that match all words in a seed fact and generates **(sentence, fact)** pairs.
- **BONIE** uses **Illinois Quantifier**'s (Roy et al., TACL 2015) internal normalizations to match quantities and units.
- A percentage threshold controls the amount of allowed difference between quantities in the sentences and the seed facts.

Open Pattern Learning

- For each **(sentence, fact)** pair, **BONIE**
  - parses the sentence.
  - replaces argument and relation words of the fact with {arg} and {rel}
  - replaces quantity or unit word with {quantity} depending on which one is at a higher level in the parse.
- Minimal path containing {arg}, {rel} and {quantity} is learned as a pattern.

Precision-Yield Curve



- **BONIE** achieves substantially larger area under the curve than **Open IE 4.2**.
- **BONIE** has **1.5x** yield and **15 point** precision gain on numerical facts over **Open IE 4.2**.

Constructing Extractions

- Arg/relation phrases are completed by expanding the extracted headwords on different dependency labels; quantity phrase is extracted by the **Illinois Quantifier**.

Relation Phrase Construction

- Whenever the relation headword is an adjective or adverb, **BONIE** uses **Wordnet** to replace the relation with its derivationally related noun form.
  e.g. (Donald Trump; old; about 70 years) -> (Donald Trump; has age of; about 70 years)
- **BONIE** uses **UnitTagger** (Sarawagi and Chakrabarti, AAAI 2014) to infer implicit relations from units in extractions.
  e.g. (James Valley; has sq kms of; 5 of fruit orchards) -> (James Valley; has area of fruit orchards; 5 sq kms)
- If a plural noun relation word appears as a unit in the quantifier, **BONIE** hypothesizes it as a count extraction and appends 'number of'.
  e.g. (Microsoft; has employees; 100,000 employees) -> (Microsoft; has number of employees; 100,000)

Experiments and Results

- **BONIE** is built over 20 million numerical sentences from **ClueWeb12**.
- **BONIE**
  - learns **21,000** seed facts.
  - bootstraps **18,500** (sentence, fact) pairs.
  - learns **7,000** new patterns.
- Comparison of precision and yield (# correct extractions) for each setting on a dataset of 2000 **ClueWeb12** numerical sentences.

Setting	Precision	Yield
NumberRule	50.00	6
Open IE 4.2	62.50	296
BONIE(seed patterns only)	85.71	72
+ learned patterns	13.88	362
+ fact filters	55.27	351
+ Yago + Wordnet expansion on facts	72.69	418
+ Relation phrase construction	77.91	448
+ Wordnet expansion on patterns	77.23	458

- **BONIE** has **1.5x** yield and **15 point** precision gain on numerical facts over Open IE .4.2.

Conclusion and Future Work

- Two-third of **BONIE**'s missed recall is because of **missing conjuncts**.
  e.g. *The retirement age for men is 65 years and 68 years for women.*
- We release **Open IE 5.0**, which improves upon **Open IE 4.2** by handling noun relations(**RelNoun 2.2**), numerical relations(**BONIE**) and **conjuncts** better.
  https://github.com/dair-iitd/OpenIE-standalone