Open Numerical Relation Extraction

Master's Thesis Presentation

Hasan Shahid Malik

Heidelberg University
Institute of Computer Science
Database Systems Research Group
er458@uni-heidelberg.de

April 24, 2020

Outline

Outline

Numerical Relations

- The area of Germany is 357,022 square kilometers.
- Carbon is a chemical element with the symbol C and an atomic number of 6.

Numerical Relations

- The area of Germany is 357,022 square kilometers.
- Carbon is a chemical element with the symbol C and an atomic number of 6.

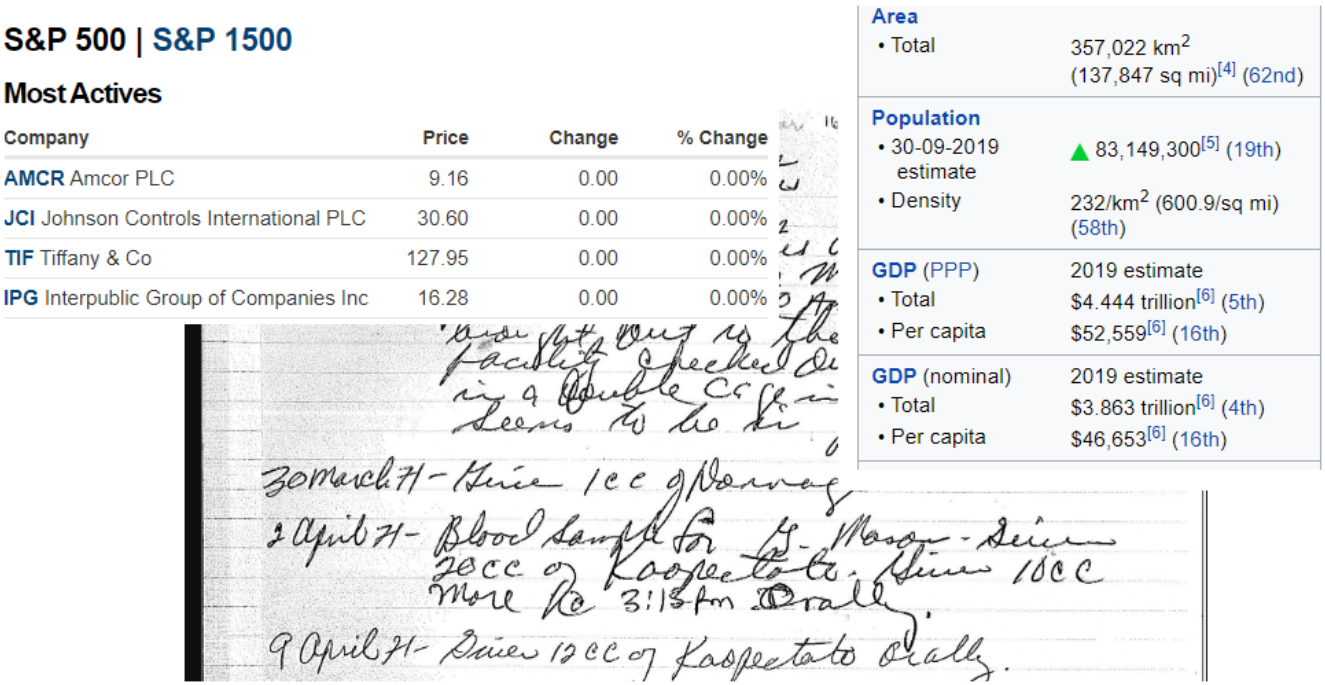
Numerical Relation Extraction (NRE)

The task of extracting semantic relationships between an entity and a quantity.

First introduced by [Madaan et al., 2016].

Applications of NRE

NRE has applications in text summerization (wikipedia infobox), database population, information extraction (EHR), etc.



S&P 500 | S&P 1500

Most Actives

Company	Price	Change	% Change
AMCR Amcor PLC	9.16	0.00	0.00%
JCI Johnson Controls International PLC	30.60	0.00	0.00%
TIF Tiffany & Co	127.95	0.00	0.00%
IPG Interpublic Group of Companies Inc	16.28	0.00	0.00%

Area	
• Total	357,022 km^2 (137,847 sq mi)[4] (62nd)
Population	
• 30-09-2019 estimate	▲ 83,149,300[5] (19th)
• Density	232/km^2 (600.9/sq mi) (58th)
GDP (PPP)	2019 estimate
• Total	$4.444 trillion[6] (5th)
• Per capita	$52,559[6] (16th)
GDP (nominal)	2019 estimate
• Total	$3.863 trillion[6] (4th)
• Per capita	$46,653[6] (16th)

30 March H- Heine 1cc g Dorvag
2 April H- Blood sample for G. Mason - Heine 20cc of Kaopectate. Heine 10cc More Rx 3:15 pm Orally.
9 April H- Heine 12cc of Kaopectate Orally.

Figure: Uses of NRE.

Motivation

Open Information Extraction

The task of extracting relation tuples without a pre-specified relation list.

Open IE methods studied predominantly on non-numeric relations: [Etzioni et al., 2008, Fader et al., 2011]

First, and only, Open NRE model proposed by [Saha et al., 2017].

Problem Statement

Non-Open NRE methods have poor scalability as they require training data per relation. [Madaan et al., 2016]

Only one Open NRE model exists: BONIE [Saha et al., 2017]

- Requires lots of training data ($>$ 27M sentences).
- Only extracts one relation per sentence.

We have nothing to compare it to. Need to explore alternative extraction techniques.

Objectives

Develop a new NRE model that:

- does not require pre-specified relations (i.e. is Open);
- does not require much/any labeled training data;
- outputs extractions for every numerical relation instance.

Aim:

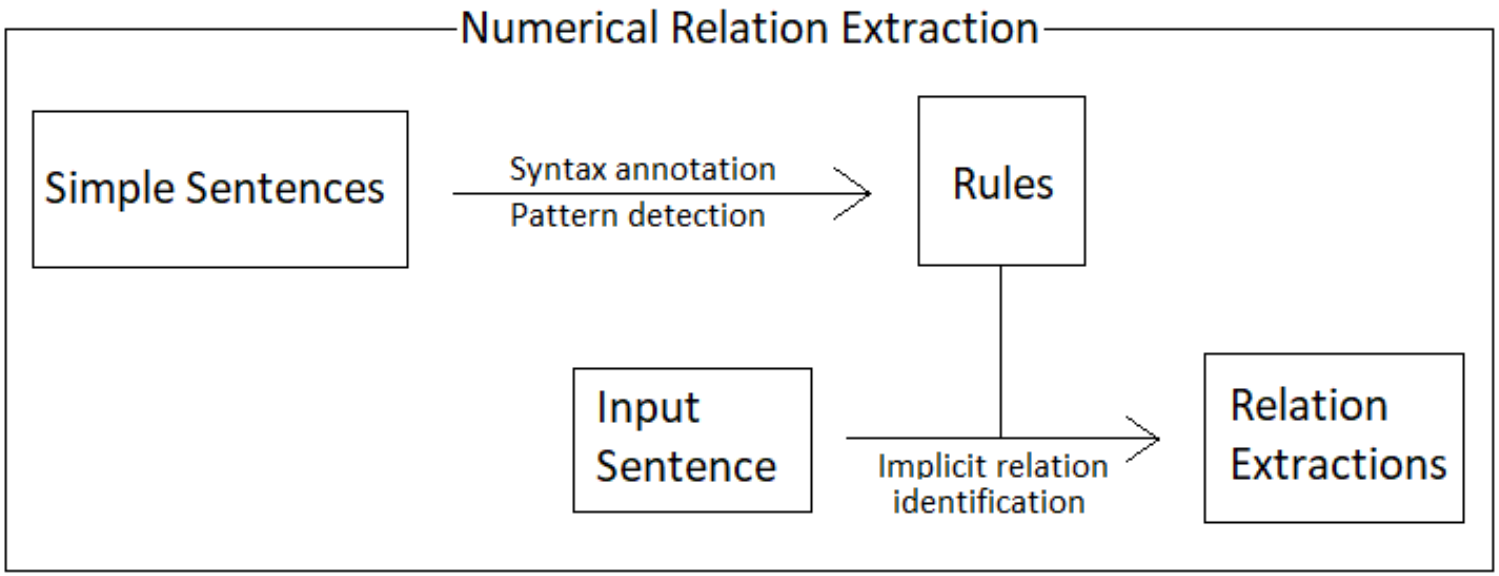
Given a sentence containing a quantity, extract a relation tuple of the form (entity, relation, verb, quantity).

Outline

Open NRE

To address the problems, we create the first rule-based Open NRE model:

1. Generalize syntax patterns in simple numerical relations to form rules.
2. Apply rules on input sentence to form extractions.
3. If unit of quantity is recognized, identify implicit relation.



Creating Rules

By studying simple numerical relations, we create 7 sets of rules that identify the entity and relation keywords based on their positions.

Germany	has	a GDP	of	$3.7 trillion
<entity>	<verb>	<relation>		<quantity>

Example ruleset:

- Entity is noun immediately preceding the verb.
- Relation is first noun preceding the quantity.

Extraction Example

Given input sentence, extract as follows:

With area of 30 m^2 , the room is spacious.

1. Identify quantity by POS tag.

Extraction Example

With	area	of	30	m^2	,	the room	is	spacious.
	NOUN		NUM	NOUN		NP	VERB	
			<quantity>					

1. Identify quantity by POS tag.
2. Identify relation-mediating verb.

Extraction Example

With area of 30 m^2 , the room is spacious.
NOUN NUM NOUN NP VERB
<quantity> <verb>

1. Identify quantity by POS tag.
2. Identify relation-mediating verb.
3. Apply rules.

Extraction Example

With	area	of	30	m^2	,	the room	is	spacious.
	NOUN		NUM	NOUN		NP	VERB	
			<quantity>				<verb>	

1. Identify quantity by POS tag.
2. Identify relation-mediating verb.
3. Entity is noun phrase immediately preceding verb.

Extraction Example

With	area	of	30	m^2	,	the room	is	spacious.
	NOUN		NUM	NOUN		NP	VERB	
			<quantity>			<entity>	<verb>	

1. Identify quantity by POS tag.
2. Identify relation-mediating verb.
3. Entity is noun phrase immediately preceding verb.
4. Relation is noun preceding quantity.

Extraction Example

With	area	of	30	m2	,	the room	is	spacious.
	NOUN		NUM	NOUN		NP	VERB	
	<relation>		<quantity>			<entity>	<verb>	

1. Identify quantity by POS tag.
2. Identify relation-mediating verb.
3. Entity is noun phrase immediately preceding verb.
4. Relation is noun preceding quantity.
5. If entity, verb or quantity is missing, extraction is skipped; otherwise, output extraction.

Extraction: (the room, area, is, $30m^2$)

Rule-Based Framework

- Rules are intuitive and easy to create.
- Model scales well w.r.t. different relations.
- However, language is versatile. Unlikely that this approach can be scaled to all possible sentence structures.

Implicit Relations

Certain numerical relations imply a relation without explicitly stating it. We use a quantity extractor [1] that identifies the implicit relation.

This bag is 5kg.
(This bag, ?, is, 5kg.)
kg implies weight
(This bag, weight, is, 5kg.)

[1] www.github.com/nielstron/quantulum3

Outline

Evaluation

Test data: 2000 sentences from ClueWeb12 [2]

- Many contain grammatical inaccuracies.
- Wide range of sources and topics.

Example sentences

- Each Qty $20.38 Pack Qty $168.03
- The Nasdaq rose 3.8 percent, its biggest weekly gain in nearly 18 months.

[2] www.lemurproject.org/clueweb12.php/

Evaluation /2

Compare extractions against BONIE [Saha et al., 2017]

- Only other Open NRE model.
- Bootstraps 6 extraction patterns over corpus to extract relation tuples, that are used to build new extraction patterns.
- Uses WordNet [3] to build relation phrase from keywords.

There is no standard evaluation metric for Open IE models. [Stanovsky and Dagan, 2016]
Therefore we use same metric that BONIE uses:

- Yield - number of correctly extracted relations.

[3] https://wordnet.princeton.edu/

Results

Model	Yield
BONIE (seed patterns only)	72
BONIE	458
Our Model	148

Table: Model yields on test data set.

Results

Model	Yield
BONIE (seed patterns only)	72
BONIE	458
Our Model	148

Table: Model yields on test data set.

Unlike BONIE, our model extracts all relations in a sentence:

"This recipe yields 6 supper servings, 12 appetizer servings."

- (This recipe, yields, 6 supper servings)
- (This recipe, yields, 12 appetizer servings)

Results /2

In some cases our model performs better at separating entity and relation:

"In 1996, Israel's GDP per capita was $17,200."

- BONIE: *(Israel's GDP per capita, was, $17,200)*
- Ours: *(Israel, GDP per capita, was, 17200 dollar)*

However, our model is prone to forming awkwardly-worded extractions:

"Sturdy handle is 17 inches high."

- BONIE: *(Sturdy handle, has height of, 17 inches)*
- Ours: *(Sturdy handle, height, is, 17 inch)*

Outline

Conclusions

- Our rule-based model eliminates need for training data, but produces fewer extractions.
- Rule-based models are prone to forming awkwardly worded extractions.
- Syntax-based methods can have difficulty detecting entity boundaries.
- Existing Open NRE models perform poorly on general text, and need further study.

Future Work

- Find ways to improve entity detection.
- Incorporate verb into relation phrase.
- Adjust implicit relation keyword conjugations for grammatically correct extractions.
- Extract numerical relations between multiple entities (e.g. "distance between X and Y is . . .").

References I

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008).
Open information extraction from the web.
Communications of the ACM, 51(12):68.

Fader, A., Soderland, S., and Etzioni, O. (2011).
Identifying relations for open information extraction.
In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Madaan, A., Mittal, A., Mausam, Ramakrishnan, G., and Sarawagi, S. (2016).
Numerical relation extraction with minimal supervision.
In *AAAI*.

References II


Saha, S., Pal, H., and Mausam (2017).
Bootstrapping for Numerical Open IE.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323, Vancouver, Canada. Association for Computational Linguistics.


Stanovsky, G. and Dagan, I. (2016).
Creating a Large Benchmark for Open Information Extraction.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, Austin, Texas. Association for Computational Linguistics.