ACL 2017 July 30-August 4, 2017 · Vancouver, Canada

Cardinal Virtues: Extracting Relation Cardinalities from Text

Paramita Mirza[1], Simon Razniewski[2], Fariz Darari[2] and Gerhard Weikum[1]

[1] Max Planck Institute for Informatics, Germany
[2] Free University of Bozen-Bolzano, Italy

max planck institut informatik

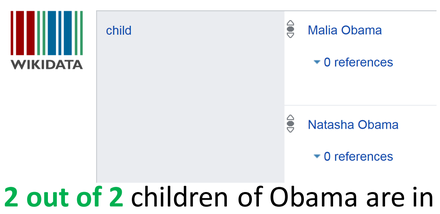
Freie Universität Bozen
Libera Università di Bolzano
Università Liedia de Bulsan

1. Overview

- IE has largely focused on answering "Who has won which award?"
- However, some facts are never fully mentioned and no IE method has perfect recall
  - Sentences like *"John lives with his spouse and 5 children on a farm in Alabama"* are much more frequent in texts.
- We focus instead on answering "How many awards has someone won?"
  - Useful for aggregate query answering, e.g., *"Who won the most awards?"*
- Contributions:
  - We introduce the problem of **Relation Cardinality Extraction**
  - We present a distant supervision method using Conditional Random Fields
  - We discuss specific challenges that set it apart from standard IE

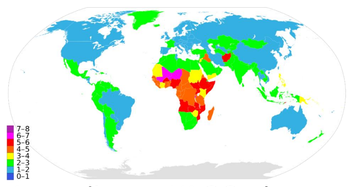
2. Motivation A: Knowledge Base (KB) curation

167 out of 199 Nobel laureates in Physics are in DBpedia ☺



2 out of 2 children of Obama are in Wikidata ☺

DBpedia contains currently only **6 out of 35** Dijkstra Prize winners ☹



According to YAGO, the average number of children per person is **0.02** ☹

KB recall is highly variant and mostly unknown ☹

*"Barack and Michelle Obama have **two** children, which are currently …."*

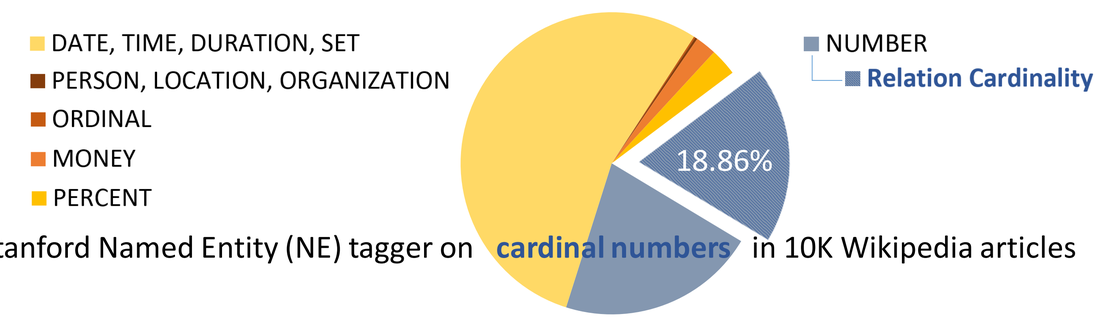
KB: 0 → Recall: 0%
KB: 1 → Recall: 50%
KB: 2 → Recall: 100%

Relation Cardinality

a mention that expresses relation cardinality
is
a cardinal number that states the number of objects that stand in a specific relation with a certain subject

*"Barack and Michelle Obama have **two** children, which are currently …."*

3. Motivation B: Disregarded by state-of-the-art (Open) IE systems



Stanford Named Entity (NE) tagger on **cardinal numbers** in 10K Wikipedia articles

Despite its frequency ☹

- Open IE (Mausam et al. 2012; Del Corro and Gemulla, 2013)
  - No way to interpret the numeric expression in the Object slot , e.g., <*Obama*, *has*, *two children*>
- KB-population IE, e.g., NELL (Mitchell et al., 2015)
  - Knows 13 relations about the number of casualties and injuries in disasters, e.g., <*Berlin2016attack*, *hasNumOfVictims*, *32*>
  - Contains only seed facts and no learned facts

4. Relation Cardinality Extraction

*"Given a well defined relation/predicate **p**, a subject **s** and a corresponding text about s, we try to estimate the **relation cardinality**, i.e., the count of <**s**, **p**, ∗> triples"*

Methodology

- Sequence labelling problem:

  Barack and Michelle Obama have two children , which are currently ….
  Barack and Michelle Obama have _num_ child , which be currently … → lemma
  O O O O O CHILD O O O O O

- Conditional Random Fields (CRF) model using CRF++ (Kudo, 2005)
  - Feature set: lemma of observed token *t*, context lemmas (windows size = 5), bigrams and trigrams containing *t*
- Distant supervision for generating training data
  - Given an <**s**, **p**> pair we identify:
    - the triple count |<**s**, **p**, ∗>| from **Wikidata** (Vrandečić and Krötzsch, 2014); and
    - candidate sentences from **English Wikipedia** article of **s**
    - candidate numbers (not labelled as TEMPORAL, MONEY or PERCENT) in each sentence (if any)
  - We generate training examples by labelling a candidate number **n** with **p** if **n** = |<**s**, **p**, ∗>|, otherwise, it is labelled as O, like the rest of non-number tokens
- Prediction
  - Having the annotated sentences by the CRF-based model,
  - **Relation cardinality** for a given <**s**, **p**> pair is the candidate number labelled with **p**, which has the highest confidence score (i.e., marginal probability of a token labelled as such, resulting from forward-backward inference)

Experiments

- Evaluation on manually annotated randomly sampled subjects for 4 Wikidata properties: 20 (*has part*), 100 (*contains admin.*) and 200 (*child* and *spouse*)
  - baseline: randomly select a number from a pool of numbers in text
  - only nummod: consider only candidate numbers that modify a noun

p	#s train	baseline	vanilla			only nummod		
		P	P	R	F1	P	R	F1
has part (creative work series)	261	.050	.333	.316	.324	.353	.316	.333
contains admin	18,000	.034	.390	.188	.254	.548	.200	.293
spouse	45,917	0	.014	.011	.013	.028	.017	.021
child	35,057	.112	.151	.129	.139	.320	.219	.260
child (manual ground truth)	6,408		.374	.309	.338	.452	.315	.317

5. Challenges in Relation Cardinality Extraction

Quality of Training Data

- Distant supervision from highly incomplete KB
  - e.g., manual annotation on *child* evaluation set → Wikidata is only ±50% accurate.
  - Unlike in classical IE, missing ground truth may lead to false positives as well.
- Possible approaches:
  - **Filtering ground truth** → consider only popular entities for training.
  - **Incompleteness-resilient distant supervision** → label all numbers equal or higher than the KB count as positive examples.

Compositionality

- *"They have two sons and one daughter together; he has four children from his first wife."*
  - 16% of false positives in extracting *child* cardinalities
- Possible approaches:
  - **Aggregating numbers** → in training data generation, label a sequence of numbers as correct cardinalities if the sum is equal to the KB count; in prediction step, sum up all consecutive cardinalities.
  - **Learning composition rules** → e.g., children are composed of sons and daughters.

Linguistic Variance

- Ordinals are quite common to express lower bounds, e.g., *John's first wife, Mary, …".*
- Relation cardinalities are sometimes expressed with non-numerals, e.g., *"He never married", "They have a daughter together", "The book is a trilogy".*
- Possible approaches:
  - **Translation to numbers** → translate certain kinds of negation and indefinite articles into expressions containing 0 and 1.
  - **Word similarity with cardinals** → consider words bear high similarity with cardinal numbers, possibly in other language such as Latin or Greek.

Further Reading

- *Predicting Completeness in Knowledge Bases*, Luis Galárraga, Simon Razniewski, Antoine Amarilli, Fabian M. Suchanek, WSDM, Cambridge, UK, 2017
- *Expanding Wikidata's Parenthood Information by 178%, or How To Mine Relation Cardinalities*, Paramita Mirza, Simon Razniewski, Werner Nutt, ISWC Poster, Osaka, Japan, 2016
- *But What Do We Actually Know?*, Simon Razniewski, Fabian Suchanek, Werner Nutt, AKBC workshop at NAACL, San Diego, USA, 2016
- *Identifying the Extent of Completeness of Query Answers over Partially Complete Databases*, Simon Razniewski, Flip Korn, Werner Nutt, Divesh Srivastava, SIGMOD, Melbourne, Australia, 2015
- A tool for crowdsourced completeness annotations for Wikidata: http://cool-wd.inf.unibz.it/

Acknowledgment
This work has been partially supported by the projects "TCFR - The Call for Recall", funded by the Free University of Bozen-Bolzano.