Expanding Wikidata's Parenthood Information by 178%, or How To Mine Relation Cardinalities

Paramita Mirza[1], Simon Razniewski[2] and Werner Nutt[2]

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

Abstract. While automated knowledge base construction so far has largely focused on fully qualified facts, e.g. ⟨*Obama, hasChild, Malia*⟩, the Web contains also extensive amounts of cardinality information, such as that someone has two children without giving their names. In this paper we argue that the extraction of such information could substantially increase the scope of knowledge bases. For the sample of the *hasChild* relation in Wikidata, we show that simple regular-expression based extraction from Wikipedia can increase the size of the relation by 178%. We also show how such cardinality information can be used to estimate the recall of knowledge bases.

1 Introduction

General-purpose knowledge bases (KBs) such as Wikidata [5], YAGO [4] or the Google Knowledge Vault [1] try to capture as much information about the world as possible. While they usually have a high precision (for instance >95% for YAGO), their recall is much lower (e.g., only 6 out of 35 Dijkstra prize winners are in DBpedia, or only about 0.02% of all living people are currently in Wikidata), and in general hard to assess [3,2]. While extraction techniques are continually improving, there exist some fundamental barriers to high recall, as many facts, such as the favourite dishes of the authors of this paper, are just not present on the web, so better extraction techniques will not help.

But there is some hope. For a substantial set of topics, natural language texts contain cardinality information, such as "John wrote two books", or "Mary has three children"). While such cardinality information does not allow to recreate fully qualified facts, it still carries interesting information, and can be useful for instance for directing KB authors towards incomplete parts, for informing data consumers about missing data, or for improving the precision of query results (e.g., "Give me the average number of children per person").

Most common data models support partially qualified information, RDF for instance via blank nodes, SQL via nulls, and OWL via cardinality constraints. Cardinality information can also be found in Wikidata, which has a property called *number of children (P1971)*. It is scarcely used so far however, i.e., only 0.21% of *humans* in Wikidata have it (6,740 in total).

In this paper we exemplify the extraction and use of cardinality information for the *hasChild* relation in Wikidata. Our technical contribution is threefold:

1. We show that cardinality assertions exists numerously in Wikipedia, thus confirming the motivation for data models that allow to specify cardinality constraints, blank nodes, labeled nulls, and similar.
2. We show that with simple filters, we can extract high quality cardinality assertions having >90% precision, which allow us to learn about the existence of 178% more children than there are currently in Wikidata.
3. We show how this information can be used to assess the recall of existing KBs, finding for instance that *child* data is almost 10 times more complete for *actors* (2.42%) than for *association football players* (0.25%).

Our extracted cardinality assertions are available online.[1]

2 Extracting Cardinality Information

In natural language texts, cardinality information for children is expressed by phrases such as:

1. *The couple had 6 children.*
2. *He never had any children.*
3. *They are the parent of three beautiful daughters.*
4. *Barnes has 2 sons and one young daughter.*

In this work, we use surface patterns via regular expressions to extract cardinalities. We manually constructed 30 patterns to find such sentences and to determine the total number of children according to the cardinal numbers found in the sentences. Our method is able to resolve, for instance, that according to Sentence 2 the total number of children is zero, or three for Sentence 4.

A major challenge in information extraction is entity resolution. We avoid this challenge by working only on biographical articles in Wikipedia, and assuming that children cardinalities mentioned in texts refer to the number of children of the person the article is about. To reduce the number of incorrect assertions that may result from this, we propose two filters:

1. *1-statement filter.* This filter removes all articles that contain more than one cardinality statement. The intuition is that even if cardinalities of multiple statements match, it is hard to decide whether one of the statements is just wrong or redundant, or whether they should be summed (frequently, articles would describe children counts from different marriages in separate sentences).
2. *75%-shortest filter.* This filter removes the 25% longest articles, based on the observation that longer articles frequently contain children information of parents or other relatives ("His son John is a successful lawyer that lives with his wife and two children in New Hampshire").

[1] `http://paramitamirza.com/other/cardinality-statements/`

Table 1. Precision on 50 samples (gold) and the *number of children* property (silver).

	#statements	gold standard			silver standard		
		#stmts	#correct	prec.	#stmts	#correct	prec.
all statements	123,885	50	43	.860	3,156	2,626	.832
1-statement filter	112,654	45	41	.911	2,815	2,496	.887
75%-shortest filter	92,914	37	34	.919	1,612	1,416	.878
both filters	86,227	35	33	.943	1,506	1,366	.907

Evaluation. We evaluate the precision of our extraction in two ways: (i) manual evaluation on 50 random phrases expressing children cardinalities (gold standard) and (ii) comparison of the extracted cardinality statements with the values of the *number of children* property (silver standard). Table 1 shows the evaluation results in which our unfiltered extraction achieves 86.0% and 83.2% precision for gold and silver standard, respectively, for a total of 123,885 extracted assertions. After applying both filters, 86,227 assertions remain, with a precision of 94.3% and 90.7%, respectively. Note that the lower precision on the silver standard likely comes from the fact that the *number of children* property itself can be outdated or may contain errors. For 2,289 out of these 86,227 persons, all children are already contained in Wikidata. The remaining 83,938 persons are missing 287,153 children, 178% more than the number of *child* facts currently contained in Wikidata.

3 Using Cardinality Information to Estimate KB Recall

Given the cardinality statements that we extracted, children information is complete for 0.7‰ of the 3.14 million humans currently contained in Wikidata (which however, in turn, are only about 0.03‰ of all the people that ever lived[2]). For those humans for which we could extract a cardinality assertion, in turn, 2.65% have complete children in Wikidata. As it is interesting to know in which parts knowledge bases are more complete, in the following we do a simple analysis based on dead/alive and occupations of persons.

Dead vs Alive. Cardinality statements extracted from articles are more likely to be found in articles of persons that are dead (3.81%), than for those that are alive (1.99%). Similarly, for those having a cardinality assertion, the *child* relation is more likely to be complete for dead (1.72%) than for living humans (0.88%). One might conjecture that for dead people, it is easier to consolidate data.

Occupations. Based on 20 most frequent occupations in Wikidata, we found that *judges* (8.22%), *lawyers* (7.93%), and *politicians* (5.11%) are the top occupations

[2] `https://en.wikipedia.org/wiki/World_population#Number_of_humans_who_have_ever_lived`

with cardinality information available in their Wikipedia articles; compared with sportsmen, e.g., *association football player* (0.51%), *athletics competitor* (1.27%), *ice hockey player* (1.10%) that seldom have such information. In turn, comparing actual *child* facts in Wikidata with extracted cardinality information, we find that matches most frequently happen for showbiz-related professions such as *actor* (2.42%) or *film director* (2.79%), and again least frequent for sport players, e.g. *ice hockey player* (0.0%) or *baseball player* (0.13%).

4 Outlook

Given available numerous cardinality information for the *child* relation in Wikipedia, we have presented a simple method to extract high quality cardinality assertions, which we then used to assess the completeness of the relation.

A challenge in broadening this work is that for weakly-defined relations such as *hobby* or *profession*, cardinality is difficult to assert. We plan to focus next on other well quantifiable relations such as as *sibling* ("*He has 3 older brothers*"), *graduatedFrom* ("*She holds a PhD in Chemistry*"), and in particular intellectual work ("*He has written two books, she composed 5 operas, he directed 12 movies*").

There are several ways to improve the quantity and quality of extracted cardinality statements. Cardinality information found in Wikipedias in other languages could and further pattern engineering could be used both for retrieving more statements, or for improving the precision. For retrieving more statements, one could also drop our restriction to biographical Wikipedia articles and our filters. This may decrease precision though, as co-reference resolution for entities expressed via pronouns ("They"), incomplete names ("Barnes"), or generic nouns ("the couple") is still a challenging NLP task.

Acknowledgment

This work has been partially supported by the projects "MAGIC", funded by the province of Bozen-Bolzano, and "The Quest to Know What We Know", funded by the Free University of Bozen-Bolzano.

References

1. X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.
2. L. Galàrraga, S. Razniewski, A. Amarilli, and F. M. Suchanek. Predicting completeness in knowledge bases. *Manuscript*, 2016. Available at `http://luisgalarraga.de/manuscripts`.
3. S. Razniewski, F. M. Suchanek, and W. Nutt. But what do we actually know? *AKBC*, 2016.
4. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, 2007.
5. D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014.