Cardinality Extraction from Text for Ontology Learning

Monika Jain
monikaja@iiitd.ac.in
Knowledgeable Computing and Reasoning Lab, IIIT-Delhi, India

Paramita Mirza
paramita@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbrücken, Germany

Raghava Mutharaju
raghava.mutharaju@iiitd.ac.in
Knowledgeable Computing and Reasoning Lab, IIIT-Delhi, India

1 INTRODUCTION

An ontology is a formal description of knowledge as a set of concepts and the relationships that hold between them. Ontologies are used in a variety of domains and in applications that involve question answering, recommendations, data integration, inferencing, etc. In order to make sense of large amount of unstructured data such as text, it needs to be in a structured form that is suitable for machines to work with. Ontologies can help in this regard. Building an ontology manually is a time consuming process. Domain experts are required to verify the vocabulary used and the correctness of the ontology. These domain experts are in some cases not available or very expensive to engage with for the length of the time required to build the ontology. Another alternative is to mine the existing human knowledge in the form of books, documents, scholarly papers, etc., and build an ontology automatically so that this knowledge is now machine processable. Building an ontology automatically is referred to as *ontology learning* [1].

There have been several efforts in the past [3] on learning ontologies from text. Statistics, linguistics, and logic based techniques were used to extract concepts and relationships from text. But almost all the existing ontology learning systems generate only shallow subclass relationships between the concepts. In this work, we look at cardinality relations between the concepts. We will, in particular, focus on extracting three types of cardinality relations, at least (minimum), at most (maximum) and exact. Cardinalities play an important role in several domains and it is generally hard to extract relations involving cardinalities from text. In the healthcare domain, information involving cardinalities is crucial and can have a major impact on the type of automated recommendations and predictions that can be made based on the data. An example of maximum cardinality from the neonatal domain is as follows - *Preterm babies have a maximum gestational age of 26 weeks.* Here the concepts, preterm babies and duration are linked to each other using a maximum cardinality relation on gestational age. Our objective is to extract this type of cardinality relations in the neonatal domain from textual sources such as Neofax [4]. Neofax contains information for around 200 medications for neonates, along with description of dosage, frequency of administering the medications, medication incompatibilities, etc. An ontology with this information will be very useful for a medication recommendation application. Along with recommending medications, it can also help the doctors to identify medications that are incompatible with each other.

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India

ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371223

2 APPROACH

We are following the lead from [2], where simple cardinality relations (not involving maximum, minimum, and exact) were extracted and added to an existing Knowledge Graph (which has only instance data). Following steps are involved in extracting cardinalities from the text and adding them to the ontology.

(1) Pre-process the Neofax document to remove unnecessary whitespace, references, and tables.
(2) Extract sentences which contain numerical values and keywords such as at least, at most, equal to, and their synonyms greater than, less than, etc., from Neofax. These sentences contain instances (particular medication such as Amikacin) and the cardinality relation between them.
(3) Create templates from the extracted sentences. Most of the sentences in Neofax has similar structure while describing particular aspects (such as dosage) of different medications. These templates can be used to create labelled data for training machine learning models.
(4) These models can be used to extract the necessary information from the sentences that do not follow the template structure. They can also be used on other related textual data such as PubMed articles. Further, We plan to do a user study to evaluate the extracted cardinality relations.
(5) From among these extracted cardinality relations, group the ones that have similar cardinality characteristics for certain type of medications. We can use the information (features) of different medication categories such as antibiotics, vaccinations, etc.
(6) After grouping them, the cardinality information can be elevated to the level of the schema (ontology). For example, we can add information such as *antibiotics can treat at least one infection.*

We are currently implementing these steps. After the ontology is ready, neonatologists will give us the feedback on the ontology and also on the medication recommendation application developed using this ontology.

REFERENCES

[1] Jens Lehmann and Johanna Voelker. 2014. An Introduction to Ontology Learning. In *Perspectives on Ontology Learning*. IOS Press, ix–xvi.
[2] Paramita Mirza, Simon Razniewski, Fariz Darari, and Gerhard Weikum. 2018. Enriching Knowledge Bases with Counting Quantifiers. In *17th International Semantic Web Conference (ISWC), October 8-12, 2018*. Springer, 179–197.
[3] Wilson Wong, Wei Liu, and Mohammed Bennamoun. 2012. Ontology Learning from Text: A Look Back and into the Future. *ACM Computing Surveys* 44, 4, Article 20 (September 2012), 36 pages.
[4] MD Young, Thomas E. 1998. *Neofax*. Eleventh edition. Raleigh, NC : Acorn Publishing, 1998. https://search.library.wisc.edu/catalog/999831778802121