TECHNISCHE UNIVERSITÄT DRESDEN

FACULTY OF COMPUTER SCIENCE
INTERNATIONAL CENTER FOR COMPUTATIONAL LOGIC

Master Thesis

Master Computational Logic

Translating Natural Language to SPARQL

Xiaoyu Yin
(Born 13. June 1994 in Zhumadian, Mat.-No.: 4572954)

Supervisor: Dr. Dagmar Gromann

Dresden, September 12, 2018

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die von mir am heutigen Tag dem Prüfungsausschuss der Fakultät Informatik eingereichte Arbeit zum Thema:

Translating Natural Language to SPARQL

vollkommen selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Dresden, den September 12, 2018

Xiaoyu Yin

Abstract

English abstract here

Contents

1 Introduction

This chapter provides an introduction about the motivation of this thesis in section 1.1. Moreover, an outline is listed in section 1.2.

1.1 Motivation

The World Wide Web is quickly evolving nowadays and has contributed to an enormous change of human lives in various areas. However, due to the design principles in early stages, a majority of the documents on the web serves as the content for human reading, thus lacking the description of a shared semantics for the computers to handle. The Semantic Web is the concept of a Web containing data and information that can be manipulated by machines automatically [SHBL06]. In order to help achieve the potential of the current web, a series of technologies including Resource Description Framework (RDF) and Ontology, have been introduced. With the help of these tools, a increasing number of documents containing uniform organized data have been gathered on the web. One notable example of this is Linked Dataset such as DBpedia [ABK+07].

SPARQL is a language designed to query and manipulate the information sources contained in an RDF store or online RDF graph content, and is by far the recommended standard [HS13]. Though the data queried by SPARQL is made for publicity and openness, the use of it has yet been spreaded out of a group of experts with prior knowledge specific to certain domains. The root of this problem is the gap between the natural language used by non-experts and the domain-specific query language specified in different syntax, semantic and vocabulary. The motivation of this thesis is to bridge this gap by investigating into the past and future methods, notably in the field of neural machine translation to fulfill the task of translating natural language to the expressions in SPARQL, and make comparisons between different methodology.

1.2 Thesis Outline

Chapter 2 presents the notion of Semantic Web and its corresponding technologies, research in the subject of neural machine translation under the area of deep learning, and the past work closely related to this thesis.

(TBD) Chapter 3 describes the research method used in this thesis.

Chapter 4 shows the experiments carried out to investigate better neural network models on the task of translating natural language to SPARQL, the datasets applied, and the corresponding results in textual and tablular form.

Chapter 5 depicts the analysis derived from the experiment results exhibited in chapter 4.

Chapter 6 brings a summary in general and points out potential directions for the future work.

2 Background

This chapter introduces the background knowledge involved in this thesis. Semantic Web Technologies is briefly introduced in section 2.1, including the notion of Linked Data in section 2.1.1 and SPARQL in section 2.1.2. Section 2.2 introduces the field of neural machine translation. Related research is discussed in section 2.3.

2.1 Semantic Web Technologies

The World Wide Web has changed the livings of people dramatically. It enables people from all over the world to browse, share, and communicate large amount of information in an unprecedented speed. This communication is based on the exchange of distributively stored documents of different kinds and formats. For a client-side user, the most common way of establishing such communication is by entering keywords, phrases, or sentences into a chosen search engine, and retrieving desired information from the result list of websites.

2.1.1 Linked Data

2.1.2 SPARQL

2.2 Neural Machine Translation

2.2.1 Sequence to Sequence Learning

2.2.2 Recurrent Neural Network

2.2.3 Long Short-Term Memory

2.3 Related Work

The primary focus of the investigation in this thesis is the neural network models that can be used to map natural language statements to SPARQL expressions. Despite that such models usually just perform the role of sequence to sequence learning, the specialty of SPARQL as a structured language with strictly defined syntax and vocabulary often lead to highly different experiment results compared to the common machine translation tasks where the source and target sequence are both unstructured. Therefore, we only consider all the research which deployed machine learning methods to map unstructured sequences to structured sequences as the most related work.

[CXY^{+}17] proposed an enhanced encoder-decoder framework for the task of translating natural language to SQL, a similar query language with SPARQL but targeting structured databases instead of knowledge bases. They used not only bleu, but also query accuracy, tuple recall, and tuple precision for measuring the quality of output queries, and achieved good results.

[SMM^{+}18, SMV^{+}18] proposed a generator-learner-interpreter architecture, namely Neural SPARQL Machines to translate any natural language expression to encoded forms of SPARQL queries. They designed templates with variables that can be filled with instances from certain kinds of concepts in target knowledge base and generated pairs of natural language expression and SPARQL query accordingly. After encoding operators, brackets, and URIs contained in original SPARQL queries, the pairs were fed into a sequence2sequence learner model as the training data. The model was able to predict on unseen natural language sentence, and generate encoding sequence of SPARQL for interpreter to decode.

3 Methodology

3.1 Model

3.2 Evaluation

4 Experiments

4.1 Implementation

4.2 Dataset

4.3 Runtime Details

4.4 Results

5 Analysis

6 Conclusion

Bibliography

[ABK+07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a Web of open data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4825 LNCS, pages 722–735. Springer, Berlin, Heidelberg, nov 2007.

[CXY+17] Ruichu Cai, Boyan Xu, Xiaoyan Yang, Zhenjie Zhang, Zijian Li, and Zhihao Liang. An Encoder-Decoder Framework Translating Natural Language to Database Queries. 2017.

[HS13] Steve Harris and Andy Seaborne. SPARQL 1.1 Query Language. Technical report, 2013.

[SHBL06] Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee. The semantic web revisited, 2006.

[SMM+18] Tommaso Soru, Edgard Marx, Diego Moussallem, Gustavo Publio, André Valdestilhas, Diego Esteves, and Ciro Baron Neto. SPARQL as a foreign language. In *CEUR Workshop Proceedings*, volume 2044, aug 2018.

[SMV+18] Tommaso Soru, Edgard Marx, André Valdestilhas, Diego Esteves, Diego Moussallem, and Gustavo Publio. Neural Machine Translation for Query Construction and Composition. 2018.

List of Figures

List of Tables

Acknowledgments

Acknowledgments here blablabla

Copyright Information

Copyright information here