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 13, 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 13, 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 now become a huge network contaning various kinds of resources for billions of users to interact with. A majority of the documents on the web are formatted texts in Hypertext Markup Language (HTML), serving as the content that can be rendered on computer browsers for humans to read and understand. In order to find a resource satisfying specific needs, a web user normally needs to rely on the help of search engines to retrieve and filter results from innumerable documents on the web. Even so, it takes quite more time for humans to distinguish useful materials from others than machines that can scan large number of files almost simultaneously. However, finishing such kind of tasks requires the capability of understanding the documents to be scanned and the traditional file format like HTML has made it difficult for machines to do so.

The Semantic Web is the concept of a Web where data and information can be manipulated by machines automatically [SHBL06]. There has been a set of standards for altering the current web to be more machine-readable and processable for automatic machine agents. In order to help achieve the potential of the current web, a series of relevant technologies including mainly Resource Description Framework (RDF) [CWL14] and Ontology have been introduced. With the help of these tools, an increasing number of documents containing uniform organized data have been published conveniently on the web. One notable example of this is cross-domain Linked Datasets such as DBpedia [ABK$^+$07]. DBpedia contains RDF documents that represents knowledge information extracted from Wikipedia websites, and all the documents with links to other datasets on the web consititute an interlinked ontology model. DBpedia also provides an open API for users to ask complicated queries against those documents.

SPARQL is a language designed for humans 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]. SPARQL has already been widely supported by large open datasets like DBpedia. 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 some certain domain. For example, if a user wants to ask for a list of books belonging to some certain category, he or she needs to have prior knowledge about the concepts and relations involved in describing books in RDF. The root of this problem is the gap between the natural language used by non-experts and the query language that consists of unique syntax, semantic and domain-specific vocabulary.

The motivation of this thesis is to bridge this gap of natural language and SPARQL. While the natural language and SPARQL are both able to be represented as sequences of pre-defined tokens, this can be categorized as a translation problem. In rencent years, the application of neural networks in machine translation has achieved great improvement on translation results than previously applied statistical and phrase-based methods [MWN17]. Therefore, we think of altering neural network models that have proven great performance in machine translation to apply on the task of translating natural language to the expressions written in SPARQL. We look into the effects of such alteration by conducting several experiments, and making parallel comparisons between different models with varying network configurations.

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 provides an outlook 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 2.1.2 and SPARQL in section 2.1.3. 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 at 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 RDF

2.1.2 Linked Data

2.1.3 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

6.1 Summary

6.2 Outlook

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.

[CWL14] Richard Cyganiak, David Wood, and Markus Lanthaler. RDF 1.1 Concepts and Abstract Syntax, 2014.

[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.

[MWN17] Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. Machine Translation Using Semantic Web Technologies: A Survey. 2017.

[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