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 18, 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 18, 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 gives a summary in general and provides an outlook for the future work.

2 Background

This chapter gives an introduction to the background technologies and subfields involved in this thesis. First, Semantic Web Technologies is briefly introduced in section 2.1, including the notion of RDF in section 2.1.1 Linked Data in section 2.1.2 and SPARQL in section 2.1.3. Second, section 2.2 describes the notion of neural machine translation and involved models and components. Finally, related research is discussed in section 2.3.

2.1 Semantic Web Technologies

The original Web consisted largely of documents made up of hypertexts for rendering on the browsers, the meanings of the web page are not well conveyed, thus being difficult for computers to analyze, or users to make higher-level searches. As Berners-Lee et al. stated in [BLHL01], the Semantic Web is not an individual web separate from the current one but an extension. In the Semantic Web, there is an important functionality that the machines are able to process the data and information automatically and even understand the data. The semantics of the web pages are well-encoded and displayed to the software agents owned by users or corporates to provide meaningful services. In the Semantic Web, the agents from different sources, namely producers and consumers, are able to communicate with each other by exchanging an Ontology which typically contains a taxonomy and a set of inference rules. With the ontology, the computers can define classes, subclasses, and relations among entites on the Web and perform automated reasoning like they "understand" the information [BLHL01].

The World Wide Web has linked more than 10 billion websites, and the useful contents can be served almost instantaneously to the users through search engines. Meanwhile, it is evolving from the web of documents for humans to read to a web of data and information derived from shared semantics. There are various types of programs and intelligent agents around the Web handcrafted for particular tasks, however, they usually possess little ability to deal with heterogeneous information kinds [SHBL06]. There is also a growing need for the integration of data and inforamtion, especially in areas that demand heterogeneous and diverse datasets originating from separate subfields [SHBL06]. A typical use case scenario of the Semantic Web is shown in figure 2.1.

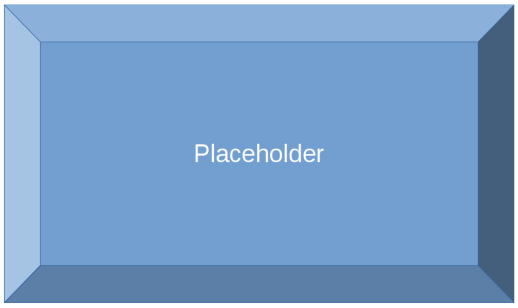

Figure 2.1: A typical use case scenario of the Semantic Web

A set of technologies are already here to provide a preliminary environment for transforming the current Web into the Semantic Web. The figure 2.2 shows an illustration of the Semantic Web technology stack, where the language in each layer is dependent on the layers below it. These languages have provided a foundation for allowing shared semantics to be integrated into the current documents on the Web, and data to be connected in a more explicit and standardized way. Resource Description Framework (RDF) [CWL14], a language located at a lower layer, has provided a foundation for the standardization of the formats of common data. SPARQL, on the other end, is a query language that can be utilized to search and manipulate data in RDF format from diverse sources [HS13]. The details of RDF and SPARQL are respectively presented in section 2.1.1 and section 2.1.3.

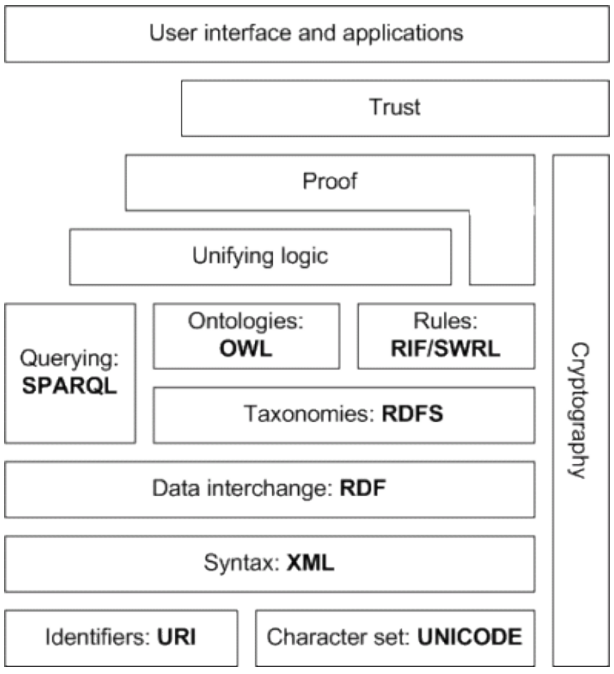

Figure 2.2: Semantic Web Technology Stack

The development of standardized technologies for the Semantic Web has promoted the integration of semantics into existing documents and linking of common data across different application domains and even regions. This enables a web of data where the data is connected with typed links, known as Linked Data [BHBL09]. It uses RDF to link arbitrary entites in the world by making typed statements, and allows complicated queries to be asked in SPARQL. The applications of Linked Data are able to work upon a global and unbound data space, whereas the conventional Web applications normally operate on top of a fixed set of data sources [BHBL09]. Further information on Linked Data is provided in section 2.1.2.

2.1.1 RDF

% What is RDF? example

% technical details perhaps

2.1.2 Linked Data

% What is Linked Data? What is the use of Linked Data?

% Which publishing tools are available for Linked Data? What's the current progress of Linked Data?

% Introduce DBpedia and its potential use, applications

2.1.3 SPARQL

% Deep introduction of SPARQL, examples, etc.

2.2 Neural Machine Translation

As a global information space, the World Wide Web contains billions of web pages written in languages from various regions of the world. The problem arises when users face the contents they request in a different language from the one they are fluent in. Translation is therefore needed here to lower the language barrier. However, it is evident that human translation does not fit the requirements due to the large quantity of web documents. One alternative solution is Machine Translation. The goal of machine translation is to transform a text from an input language to a target language with the semantic meaning of the text being preserved.

However, due to the complexity of structures, semantics and vocabularies of natural languages, translation is considered a difficulty task for machines. According to [Pop12], the errors occurred in the machine translation output are mainly classified into five base categories: inflectional error, incorrect reordering, missing word, extra word, and lexical ambiguity. There is also debate whether fully high-quality machine translation systems can be achieved [BH64]. In addition, lack of context, incomplete common sense knowledge, and ineffectiveness in translating rare words have been major issues that affect the quality of the machine translation systems [Okp14] [WSC+16].

There have been a large number of approaches in Machine Translation developed over the last years. Currently, the architectures of existing MT systems can be divided into the following categories:

- **Rule-based Machine Translation** (RBMT). RBMT systems usually generate target language text based on an intermediary linguistic representation of the source text and a large set of rules that contain morphological, syntactic, and semantic bilingual mappings. They can be further subdivided into direct, transfer-based, and interlingua-based methods. The performance of RBMT systems, to a certain extent, relies on carefully designed linguistic rules and vast amount of lexicons [MWN17].
- **Statistical Machine Translation** (SMT). SMT systems are developed on the basis of splitting a bilingual text corpora into respective source and translation text pairs. They apply Machine Learning algorithms that compute a statistical model from the corpus given and the model translates each phrase or word at a time based on a probability distribution [MWN17]. In SMT approaches, it usually requires an alignment between source sentence and several target sentences found in each text corpora and vice versa. Such methods suffer in performance when the languages involved have significantly different word orders [Okp14].
- **Example-based Machine Translation** (EBMT). EBMT utilize bilingual corpora like SMT but they translate the text by example sentences. The major limitation of EBMT systems is translation of unknown words [MWN17].
- **Neural Machine Translation** (NMT). Some argues that NMT is also a statistical approach [MWN17]. In NMT systems, they normally consist of a model based on deep neural networks to perform end-to-end translation by words or characters in the given sentence [MWN17]. During the training of the model, the system steadily learns a representation of both languages in a continuous vector space and the ability to predict a combination of words with higher probability. The approaches in this category currently achieve the state-of-the-art results on several benchmark tests. Their relevant models are the primary focus of the investigation by this thesis.
- **Hybrid Machine Translation**. Hybrid approaches essentially leverage the advantages of the

methods mentioned above to address their respective limitations and achieve better translation quality. In applications under this category, the hybridization of MT approaches are normally guided by either rule-based or corpus-based statistical systems [CJF15].

Among these categories, we focus primarily on the Neural Machine Translation methods. The development of NMT systems has gained more interests in recent years since deep neural networks have boosted extraordinary advancement in other areas of Artificial Intelligence such as computer vision [KSH12] and speech recognition [DYDA12]. NMT systems are usually superior in not needing for hand-engineering features that are one of the shortcomings of phrase-based systems [BGLL17]. However, some of the current NMT architectures have disadvantages in requiring large amount of computation and time for training the deep model [BGLL17].

So far, many different architectures have been explored in NMT and new methods are constantly beating previous models in some benchmark datasets and achieving higher efficiency in computing. Among these architectures, primary works are listed here as follows. In [SVL14, CvMG$^+$14] Sutskever et al. and Cho et al. proposed and deployed an encoder-decoder architecture that contains two models where Recurrent Neural Networks (RNN) (see section 2.2.2) were used. The encoder encodes the input into a fixed-length vector, the decoder then decodes it into a translation. The two models are jointly trained to maximize the likelihood of a target sequence based on the given source sequence. On the other hand, the performance of this architecture drops when the length of the input sentence increases. To adress this issue, Bahdanau et al. and Luong et al. presented in [BCB14, LPM15] an Attention mechanism which serves as an extension to align the encoder and decoder. The acceptance of this mechanism increased the quality of translation significantly. Furthermore, there have been other improvement strategies like bi-directional RNN, beam search, etc. Some variants like Long-Short Term Memory (LSTM) [HS97] and Gated Recurrent Unit (GRU) [CvMG$^+$14] versions of encoder-decoders have also been investigated. In the mean time, while RNN encoder-decoders comsume a lot of computation time on their sequential learning, there are architectures based on Convolutional Neural Networks (CNN) that are able to achieve parallelized computations, thus outperform the former models at a faster speed [GAG$^+$17]. What's more, Vaswani et al. proposed a self-attention model called the Transformer in [VSP$^+$17]. This model shows both quality and speed advantages and has achieves state-of-the-art results on multiple translation tasks. In the past years, some novel paradigms for NMT have also emerged like the applications of Generative Adversarial Networks (GAN) in [WXZ$^+$17, YCWX17].

2.2.1 Sequence to Sequence Learning

% What is sequence to sequence learning? What is the role of it in NMT and ML?

% Which models have been proposed in seq2seq learning?

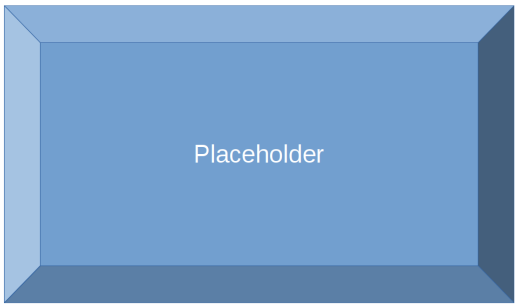

Figure 2.3: A conventional encoder-decoder framework

2.2.2 Recurrent Neural Network

% Introduction of RNN and technical details

2.2.3 Long Short-Term Memory

% Introduction of LSTM and technical details

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.

Cai et al. proposed in [CXY+17] 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 [PRWZ02], but also query accuracy, tuple recall, and tuple precision for measuring the quality of output queries, and achieved good results.

[DL16] presented a method based on encoder-decoder model with attention mechanism aimed at translating the input utterances to their logical forms with minimum domain knowledge. Apart from that, they proposed another sequence-to-tree model that has a special decoder better able to capture the hierarchical structure of logical forms. Then, they tested their model on four different datasets and evaluated the results with accuracy as the metric.

Luz et al. also used a LSTM encoder-decoder model but the purpose is to encode natural language and decode into SPARQL [LF18]. Furthermore, they employed a neural probabilistic language model to learn a word vector representation for SPARQL, and used the attention mechanism to associate a vocabulary mapping between natural language and SPARQL. For the experiment, they transformed the logical queries in the traditional Geo880 dataset into equivalent SPARQL form. In terms of evaluation, they adopted two metrics: accuracy and syntactical errors. They further compared their method with several other similar approaches [AlA15] [KBZ06] and the comparison showed that they obtained better accuracy results. However, they did not deal with the "out of vocabulary" (OOV) problem and the issue of lexical disambiguation.

In [SMM$^+$18, SMV$^+$18] Soru et al. 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 sequence to sequence 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.

[AlA15] Iyad AlAgha. Using linguistic analysis to translate arabic natural language queries to sparql. *arXiv preprint arXiv:1508.01447*, 2015.

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. sep 2014.

[BGLL17] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. Massive Exploration of Neural Machine Translation Architectures. 2017.

[BH64] Yehoshua Bar-Hillel. *Language and information: Selected essays on their theory and application*. Addison-Wesley Reading, 1964.

[BHBL09] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web, may 2001.

[CJF15] Marta R Costa-Jussa and José AR Fonollosa. Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32(1):3–10, 2015.

[CvMG+14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. jun 2014.

[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.

[DL16] Li Dong and Mirella Lapata. Language to logical form with neural attention. *arXiv preprint arXiv:1601.01280*, 2016.

[DYDA12] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2012.

[GAG+17] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional Sequence to Sequence Learning. 2017.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

[KBZ06] Esther Kaufmann, Abraham Bernstein, and Renato Zumstein. Querix: A natural language interface to query ontologies based on clarification dialogs. In *In: 5th ISWC*, pages 980–981. Springer, 2006.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[LF18] Fabiano Ferreira Luz and Marcelo Finger. Semantic Parsing Natural Language into SPARQL: Improving Target Language Representation with Neural Attention. 2018.

[LPM15] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective Approaches to Attention-based Neural Machine Translation. *EMNLP*, (September):11, 2015.

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

[Okp14] MD Okpor. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5):159, 2014.

[Pop12] Maja Popović. Class error rates for evaluation of machine translation output. *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 71–75, 2012.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. BLEU: a method for automatic evaluation of machine translation. ... *of the 40Th Annual Meeting on* ..., (July):311–318, 2002.

[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.

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112, 2014.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. jun 2017.

[WSC+16] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv e-prints*, pages 1–23, sep 2016.

[WXZ+17] Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Adversarial neural machine translation. *arXiv preprint arXiv:1704.06933*, 2017.

[YCWX17] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Improving neural machine translation with conditional sequence generative adversarial nets. *arXiv preprint arXiv:1703.04887*, 2017.

List of Figures

List of Tables

Acknowledgments

Acknowledgments here blablabla

Copyright Information