Translating Natural Language to SPARQL

Xiaoyu Yin

International Center for Computational Logic
Technische Universität Dresden

8th January 2019

Outline

Outline

- Introduction
  - Motivation
  - Background
- Methodology
  - Models
  - Datasets
  - Metrics
- Experiments
  - Experimental Setups
- Results & Discussion
- Summary & Outlook

Transformation of the Web

Motivation

Web

- Linked web pages
- Made for human to browse

Semantic Web

- Built on the existing Web
- Linked knowledge graphs and data
- For human and machines to lookup

Transformation of the Web

Motivation

Web

- Linked web pages
- Made for human to browse

Semantic Web

- Built on the existing Web
- Linked knowledge graphs and data
- For human and machines to lookup

Semantic Web Technologies

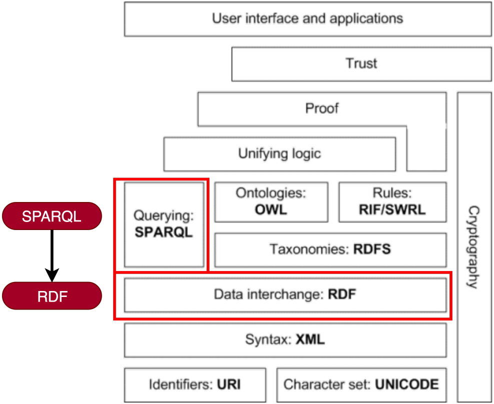


Figure: Semantic Web technology stack

Linked Data

- a notion in Semantic Web
- based on RDF
- currently only in commercial applications but has great potentials

Example

Google Knowledge Graph DBpedia

- knowledge from the Wikipedia articles in RDF files so that machines can process easily, open to everyone

SPARQL

- like SQL but for the Semantic Web and Linked Data
- structured query language
- the meaning of a SPARQL query can be usually expressed in Natural Language
- Why not Natural Language to SPARQL?

Application

- Broaden the accessibility of the Semantic Web resources
- Chatbots, service agents

Natural Language to SPARQL

- proved possible by Neural SPARQL Machine (NSpM) [SMM+18]
- no large number of tests have been done before

Our goal

Conduct a large number of tests on the possibility of translating human language to SPARQL (machine language)

Machine Translation

Background

- Rule-based Machine Translation
- Statistical Machine Translation
- Example-based Machine Translation
- **Neural Machine Translation**
- Hybrid Machine Translation

Neural Machine Translation

Given large number of training samples, use deep neural networks to perform end-to-end translation between source and target languages.

Neural Machine Translation

Background

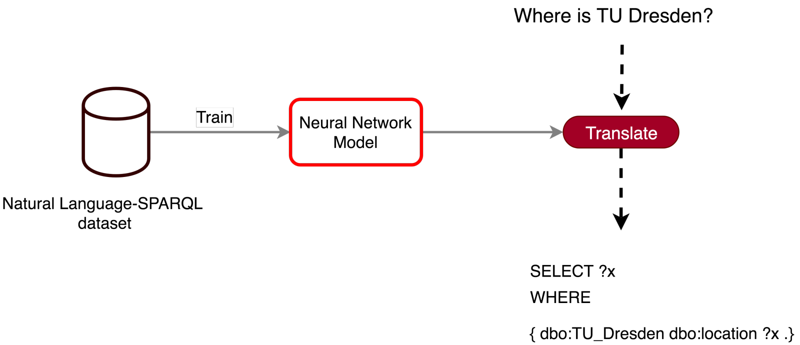
Why NMT?

- So far the best performing methods in translating between natural languages (e.g. English to German)
- Abundant choices of neural networks
- Off-the-shelf frameworks to use

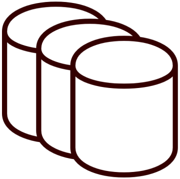
Challenges

- SPARQL is not like any Natural Language

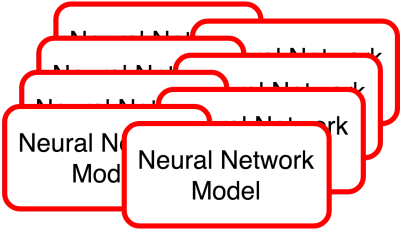
Idea


Where is TU Dresden?
Train
Neural Network Model
Translate
Natural Language-SPARQL dataset
SELECT ?x
WHERE
{ dbo:TU_Dresden dbo:location ?x .}

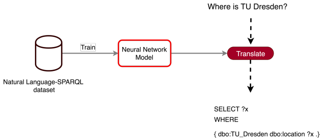
Idea



3 Datasets



8 Models



Outline

Neural Machine Translation

2013 Recurrent Neural Networks (RNN) started

2014-15 Attention mechanisms

- great enhancements to RNN

2016 Google Translate System (GNMT)

- bi-directional RNN, residual connection, etc.

2017-now Convolutional Neural Networks (CNN), Self-attention models joined

- right now state-of-the-art

Neural Machine Translation

- **2013** Recurrent Neural Networks (RNN) started
- **2014-15** Attention mechanisms
  - great enhancements to RNN
- **2016** Google Translate System (GNMT)
  - bi-directional RNN, residual connection, etc.
- **2017-now** Convolutional Neural Networks (CNN), Self-attention models joined
  - right now state-of-the-art

Neural Machine Translation

2013 Recurrent Neural Networks (RNN) started

2014-15 Attention mechanisms

- great enhancements to RNN

2016 Google Translate System (GNMT)

- bi-directional RNN, residual connection, etc.

2017-now Convolutional Neural Networks (CNN), Self-attention models joined

- right now state-of-the-art

Neural Machine Translation

2013 Recurrent Neural Networks (RNN) started

2014-15 Attention mechanisms

- great enhancements to RNN

2016 Google Translate System (GNMT)

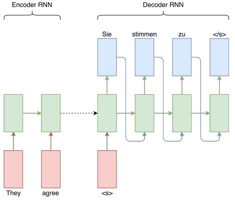
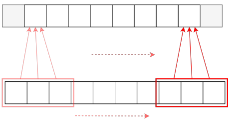
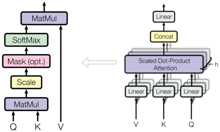
- bi-directional RNN, residual connection, etc.

2017-now Convolutional Neural Networks (CNN), Self-attention models joined

- right now state-of-the-art

Neural Machine Translation Models

- Three categories

Recurrent Neural Network (RNN) models	Convolutional Neural Network (CNN) models	Self-Attention models
		
GNMT	ConvS2S	Transformer

Google's Neural Machine Translation System (GNMT)

RNN-based

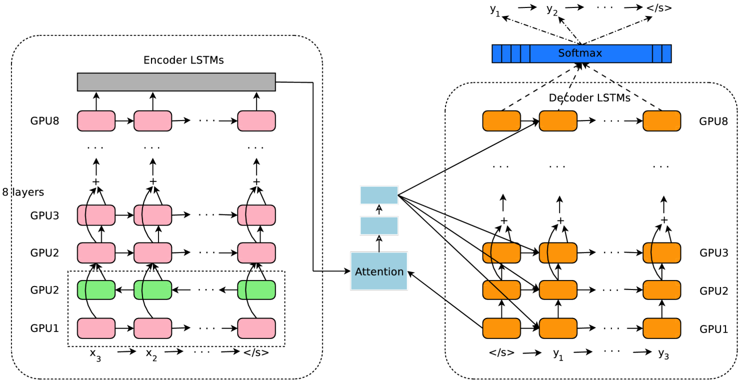


Figure: The model architecture of GNMT [WSC+16].

Convolutional Sequence-to-Sequence (ConvS2S)

CNN-based

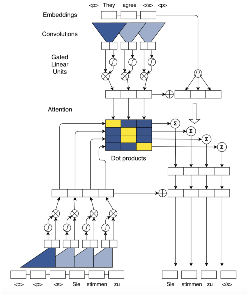


Figure: The demonstration of training the Convolutional Sequence-to-Sequence model [GAG+17]

The Transformer

Self-attention Models

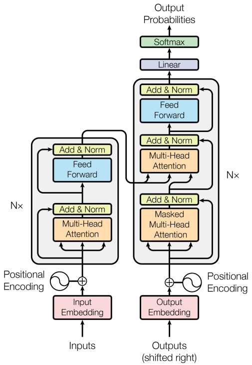


Figure: The architecture of the Transformer model [VSP+17]

Models Summary

Out of these three categories, we constructed 8 models...

- RNN-based models: **NSpM, NSpM+Att1, NSpM+Att2, LSTM_Luong, GNMT-4, GNMT-8**
- CNN-based model: **ConvS2S**
- Self-attention model: **Transformer**

- NSpM: Basic 2-layer RNN
- NSpM+Att: NSpM with attention module
- LSTM_Luong: a deep 4-layer RNN

Datasets

- The **Monument** dataset
- Largescale Complex Question Answering Dataset (**LC-QUAD**)
- DBpedia Neural Question Answering (**DBNQA**)

	Monument	LC-QUAD	DBNQA
Instance	14,788	5,000	894,499
English vocab	2,500	7,000	131,000
SPARQL vocab	2,200	5,000	244,900

Table: Sizes of three used English-SPARQL datasets

Datasets

- The **Monument** dataset
- Largescale Complex Question Answering Dataset (**LC-QUAD**)
- DBpedia Neural Question Answering (**DBNQA**)

	Monument	LC-QUAD	DBNQA
Instance	14,788	5,000	894,499
English vocab	2,500	7,000	131,000
SPARQL vocab	2,200	5,000	244,900

Table: Sizes of three used English-SPARQL datasets

Evaluation Metrics

- **Perplexity** for training phase
- **BLEU** for testing phase

Perplexity

- $1 \rightsquigarrow +\infty$
- reflects how well the model is trained (1 is best)

Evaluation Metrics

- **Perplexity** for training phase
- **BLEU** for testing phase

BLEU

- 0 $\rightsquigarrow$ 100
- reflects the quality of the generated translations compared to the reference (100 is best)
- widely used in Machine Translation tasks

Evaluation Metrics

- **Perplexity** for training phase
- **BLEU** for testing phase

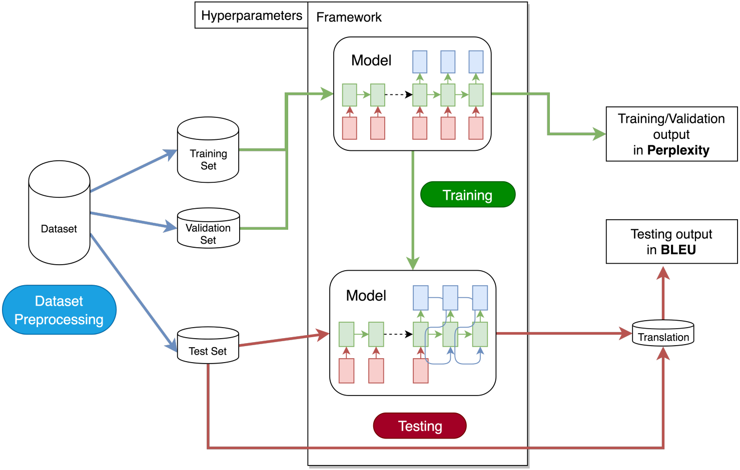
Example (BLEU)

Translation	the cat sat on the mat
Reference 1	there is a cat on the mat
Reference 2	the cat is on the mat

BLEU Score: 42

Outline

Experiment Overview



Dataset Preprocessing

Dataset Splitting

- Split the Monument, LC-QUAD, and DBNQA into 80%/10%/10% training/validation/testing set
- Split the Monument dataset further into 50%/10%/40% and 14588/100/100 (in NSpM paper)
- Leads to 5 different splits: **MonumentNSpM, Monument50, Monument80, LC-QUAD, DBNQA**

Dataset Preprocessing

SPARQL Encoding

SPARQL	`SELECT DISTINCT ?uri`
	`WHERE {`
	`<http://dbpedia.org/resource/Sam_Loyd> <http://dbpedia.org/ontology/knownFor> ?uri .`
	`<http://dbpedia.org/resource/Eric_Schiller> <http://dbpedia.org/ontology/knownFor> ?uri .`
	`}`

Encoded	select distinct var_uri where brack_open dbr_Sam_Loyd dbo_knownFor var_uri sep_dot dbr_Eric_Schiller dbo_knownFor var_uri sep_dot brack_close

Hardware

	GPU **Small**	GPU **Medium**	GPU **Large**
CPU	Intel® Xeon® CPU E5-2450 @ 2.10GHz	Intel® Xeon® CPU E5-2680 @ 2.50GHz	POWER9
RAM	24 GB	16 GB	192 GB
Cores	8	6	32
GPU	NVIDIA® Tesla® K20Xm	NVIDIA® Tesla® K80	NVIDIA® Tesla® V100-SXM2
GPU RAM	6 GB	12 GB	32 GB

Table: Three hardware configurations on High Performance Computing (HPC) server used in this thesis

Software (1/2)

Python Frameworks

- *nmt*[1] based on TensorFlow
  - Implements: **NSpM, NSpM+Att1, NSpM+Att2, GNMT-4, GNMT-8**
- *fairseq*[2] based on PyTorch
  - Implements: **LSTM_Luong, ConvS2S, Transformer**

Takes care of training, validation, and testing the models on given dataset and outputs the results and statistics

[1] https://github.com/tensorflow/nmt
[2] https://github.com/pytorch/fairseq

Software (2/2)

Operating Systems

- Linux from HPC with Python 3.6.4, TensorFlow 1.8.0, and PyTorch 0.4.1
  - Ran the training and testing jobs and saved the results
- macOS High Sierra 10.13.6 from my computer with Python 3.6.5, TensorFlow 1.8.0, PyTorch 0.4.1, and matplotlib 3.0.2
  - Preprocessed the datasets
  - Uploaded and downloaded jobs between HPC
  - Analyzed the outputs

Source code is all available on GitHub[3].

[3] `https://github.com/xiaoyuin/tntspa`

Hyperparameters

Tricky part of neural network training

1. Adopted recommended hyperparameters from each framework
2. Adjusted for the **MonumentNSpM** dataset split
3. Applied on the other splits

Outline

Results

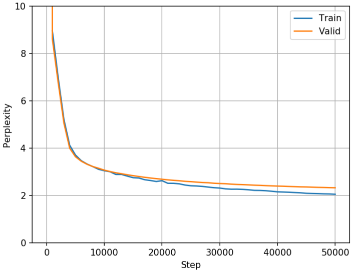
For each dataset split and model, we report...

- **Perplexity graphs** on the training and validation set
- **BEST BLEU** on the test set

In total, 5 (dataset splits) * 8 (models) = 40 perplexity graphs and 40 BLEU scores

Results

Example



Models	**Test BLEU**	Step / Epoch
NSpM	65.92	Step 50k
NSpM+Att1	89.87	Step 50k
NSpM+Att2	91.50	Step 50k
GNMT-4	69.61	Step 30k
GNMT-8	68.41	Step 30k
LSTM_Luong	77.67	Epoch 55
ConvS2S	**96.07**	Epoch 54
Transformer	68.82	Epoch 53

Dataset Comparison

Three different splits of the monument dataset did not show very big differences in results

- The **Monument** dataset is relatively simple

Serious overfit in LC-QUAD dataset experiments

- The **LC-QUAD** dataset is too small in size

- The **DBNQA** is so far the most suitable for this task

But they are all relatively simpler compared to Natural Language datasets

- 25-30 BLEU for Natural Language task, 60-100 BLEU for our task

Dataset Comparison

Three different splits of the monument dataset did not show very big differences in results

- The **Monument** dataset is relatively simple

Serious overfit in LC-QUAD dataset experiments

- The **LC-QUAD** dataset is too small in size
- The **DBNQA** is so far the most suitable for this task

But they are all relatively simpler compared to Natural Language datasets

- 25-30 BLEU for Natural Language task, 60-100 BLEU for our task

Dataset Comparison

Three different splits of the monument dataset did not show very big differences in results

- The **Monument** dataset is relatively simple

Serious overfit in LC-QUAD dataset experiments

- The **LC-QUAD** dataset is too small in size

- The **DBNQA** is so far the most suitable for this task

But they are all relatively simpler compared to Natural Language datasets

- 25-30 BLEU for Natural Language task, 60-100 BLEU for our task

Dataset Comparison

Three different splits of the monument dataset did not show very big differences in results

- The **Monument** dataset is relatively simple

Serious overfit in LC-QUAD dataset experiments

- The **LC-QUAD** dataset is too small in size

- The **DBNQA** is so far the most suitable for this task

But they are all relatively simpler compared to Natural Language datasets

- 25-30 BLEU for Natural Language task, 60-100 BLEU for our task

Dataset Comparison

Three different splits of the monument dataset did not show very big differences in results

- The **Monument** dataset is relatively simple

Serious overfit in LC-QUAD dataset experiments

- The **LC-QUAD** dataset is too small in size

- The **DBNQA** is so far the most suitable for this task

But they are all relatively simpler compared to Natural Language datasets

- 25-30 BLEU for Natural Language task, 60-100 BLEU for our task

Model Comparison

- **ConvS2S** model outperformed other models in converging speed (Perplexity curves) and translation quality (BLEU scores)
- Attention mechanisms contributed to the translation quality
- GNMT (deeper-layer model) performs relatively worse than shallower-layer models
- The Transformer model is relatively harder to train

Model Comparison

- **ConvS2S** model outperformed other models in converging speed (Perplexity curves) and translation quality (BLEU scores)
- Attention mechanisms contributed to the translation quality
- GNMT (deeper-layer model) performs relatively worse than shallower-layer models
- The Transformer model is relatively harder to train

Model Comparison

- **ConvS2S** model outperformed other models in converging speed (Perplexity curves) and translation quality (BLEU scores)
- Attention mechanisms contributed to the translation quality
- GNMT (deeper-layer model) performs relatively worse than shallower-layer models
- The Transformer model is relatively harder to train

Model Comparison

- **ConvS2S** model outperformed other models in converging speed (Perplexity curves) and translation quality (BLEU scores)
- Attention mechanisms contributed to the translation quality
- GNMT (deeper-layer model) performs relatively worse than shallower-layer models
- The Transformer model is relatively harder to train

Perplexity vs. BLEU

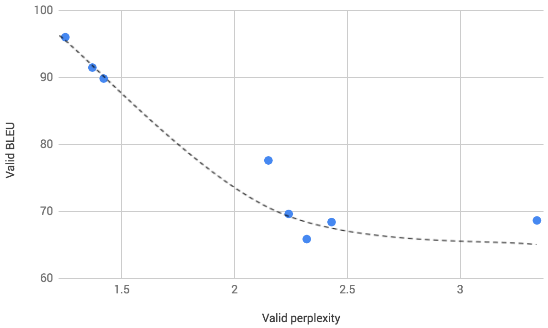


Figure: The perplexity-BLEU graph on the validation set in the DBNQA experiments

Limitations

Training

- Training hyperparameters are not tuned specifically for each model and each dataset.
- Framework differences

Testing

- BLEU is not a perfect metric for SPARQL

Outline

Summary & Outlook

- Semantic Web and SPARQL
- Neural Machine Translation Models
- Experiments
- Results and Discussion

Future Work

- a better NL-SPARQL dataset
- better metric instead of BLEU
- more hyperparameter tuning

Summary & Outlook

- Semantic Web and SPARQL
- Neural Machine Translation Models
- Experiments
- Results and Discussion

Future Work

- a better NL-SPARQL dataset
- better metric instead of BLEU
- more hyperparameter tuning

Thank you !

Reference

- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, *Convolutional Sequence to Sequence Learning*, Proc. of ICML, 2017.
- Tommaso Soru, Edgard Marx, Diego Moussallem, Gustavo Publio, André Valdestilhas, Diego Esteves, and Ciro Baron Neto, *SPARQL as a foreign language*, CEUR Workshop Proceedings, vol. 2044, aug 2018.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention Is All You Need*.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva