TriviaNLQ:
Natural Language Query for trivial questions

M2 TAL Software Project

Karolin Boczoń
Roham Roshanfekr
by
Dimitra Niaouri
Muhammad Shahzaib

Université de Lorraine

January 13, 2023

SPARQLess queries

Several issues came up while trying to run `sparqling-queries` [1]:

- broken dependencies
- complicated infrastructure, requires a database
- hard to adapt code, only works with the given dataset

Overall, it didn't work out.

Back to code generation

Turn this into Wikidata SPARQL query: "What instrument does Kermit the Frog play?".

```
SELECT ?instrument WHERE {
 wd:Q717 wdt:P361 ?instrument .
 FILTER contains(?instrument, "Kermit the Frog")
}
```

Figure: GPT-3 composed query

Looks OK at the first glance and even executes but:

- Q717 is Venezuela
- P361 is a relation ”part of”
- it keeps only the results that have ”Kermit the Frog” in its description

GPT-3 comsposed queries

- don’t execute (syntax errors, maximum limit reached)
- don’t match objects and predicates
- aren’t deterministic

ConvS2S

- Convolutional Sequence-to-Sequence model
- Based on TNTSPA [2]
- Best performing model compared to RNN-based models (NSpM, NSpM+Att1, NSpM+Att2, LSTM Luong, GNMT-4, GNMT-8) and Self-attention models (Transformer) for text-to-sparql [2]

ConvS2S

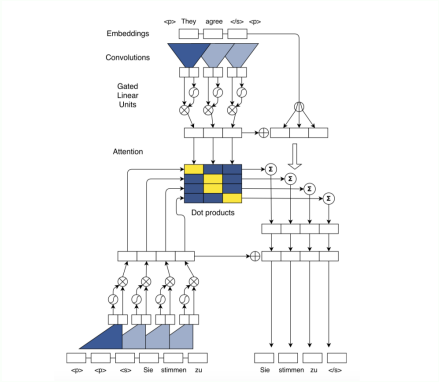

Figure: Convolutional Sequence-to-Sequence model [GAG + 17][2]

ConvS2S

Dataset

- Monument dataset
- Better results over the two other main datasets in literature (LC-QUAD and DBNQA) [2]

ConvS2S

Dataset

	Monument
Instance	14,788
English vocab	2,500
SPARQL vocab	2,200

Figure: Size of the English-SPARQL Monument dataset [2]

ConvS2S

Dataset Preprocessing

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

Figure: SPARQL Encoding [2]

ConvS2S

Dataset Splitting

- Splitting of the Monument dataset into 80%/10%/10% training/validation/testing set [2]

ConvS2S

Frameworks

- fairseq based on PyTorch (sequence modeling toolkit that allows researchers and developers to train custom models)
- implemented on ConvS2S

Hyperparameters

- Recommended hyperparameters from fairseq [2]

ConvS2S

Results

Model	Validation	Test
ConvS2S	94 | 99	95 | 96

Figure: Accuracy (in %) of syntactically correct generated SPARQL queries | F1 score)[2]

Next Steps

- Wrapper for the model
- Script for turning the encoded output into SPARQL
- Front-end
- Script for querying DBpedia
- Check the performance on targeted trivia questions

References

[1]Saparina, I., Osokin, A. (2021). Sparqling database queries from intermediate question decompositions. arXiv preprint arXiv:2109.06162.
[2]Yin, X., Gromann, D., Rudolph, S. (2019). Neural machine translating from natural language to SPARQL. Future Generation Computer Systems, 117, 510-519.