LingNLQ:
Natural Language Query for linguistics

M2 TAL Software Project

by

Karolin Boczoń
Roham Roshanfekr
Dimitra Niaouri
Muhammad Shahzaib

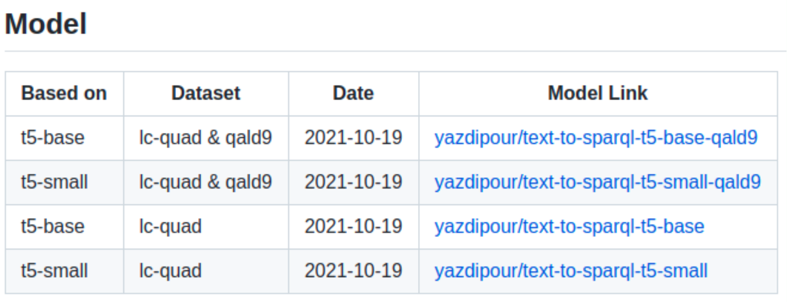
Université de Lorraine

November 7, 2022

text-to-sparql models

text-to-sparql-t5

This model is a fine-tuned version of t5-base.



Model

Based on	Dataset	Date	Model Link
t5-base	lc-quad & qald9	2021-10-19	yazdipour/text-to-sparql-t5-base-qald9
t5-small	lc-quad & qald9	2021-10-19	yazdipour/text-to-sparql-t5-small-qald9
t5-base	lc-quad	2021-10-19	yazdipour/text-to-sparql-t5-base
t5-small	lc-quad	2021-10-19	yazdipour/text-to-sparql-t5-small

Figure: Different versions of the model

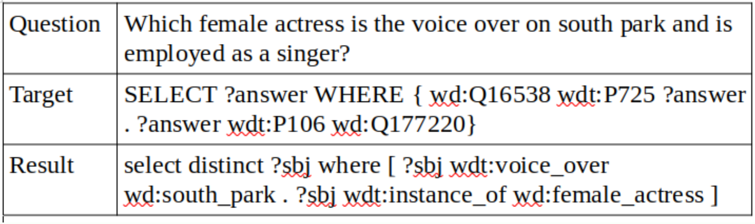
text-to-sparql-t5

Issues of the model

- The target and results’ queries were not well-formed (grammatical errors, square brackets instead of curly ones etc.)
- Poor performance

text-to-sparql-t5

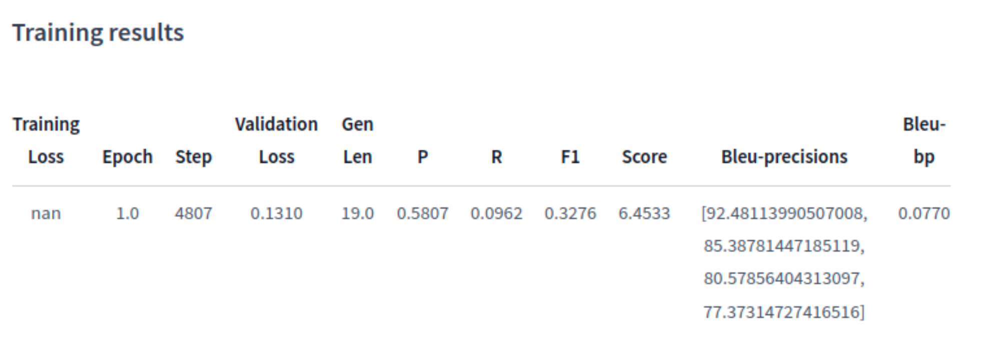
Model Architecture



Question	Which female actress is the voice over on south park and is employed as a singer?
Target	SELECT ?answer WHERE { wd:Q16538 wdt:P725 ?answer . ?answer wdt:P106 wd:Q177220}
Result	select distinct ?sbj where [?sbj wdt:voice_over wd:south_park . ?sbj wdt:instance_of wd:female_actress]

Figure: Example the Question-Target-Result architecture

text-to-sparql-t5



Training results

Training Loss	Epoch	Step	Validation Loss	Gen Len	P	R	F1	Score	Bleu-precisions	Bleu-bp
nan	1.0	4807	0.1310	19.0	0.5807	0.0962	0.3276	6.4533	[92.48113990507008, 85.38781447185119, 80.57856404313097, 77.37314727416516]	0.0770

Figure: Training results of the text-to-sparql-t5-base model

Wine Ontology

Sample program to read a NL input and generate a sparql query to query the wine ontology and get results.
Issues:

- Code written in python version 2.7
- No available results

Quepy

Python framework to transform natural language questions to queries in a database query language.

- easily customized to different kinds of questions in NL and database queries
- support for Sparql and MQL query languages

Issue: Code written in python version 2.

MK-SQuIT and NeMo

Creates datasets to train machine translation systems to convert natural language questions into queries.

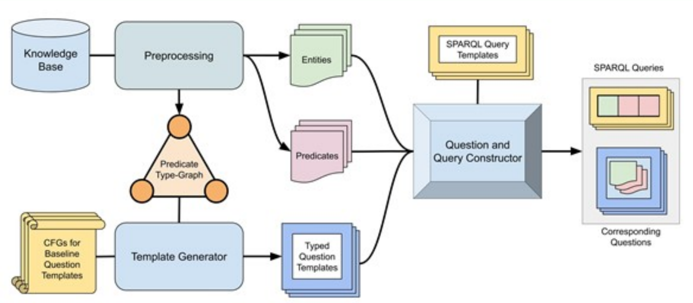
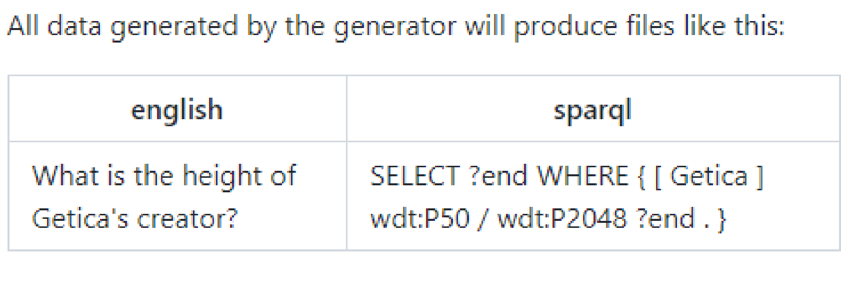


Figure: MK-SQuIT generation pipeline.

MK-SQuIT and NeMo

Data Format



All data generated by the generator will produce files like this:

english	sparql
What is the height of Getica's creator?	SELECT ?end WHERE { [Getica] wdt:P50 / wdt:P2048 ?end . }

LYMBA

Model for creating a knowledge base from text and converting text to SPARQL for widespread usage.

- question sent through the Lymba pipeline
- system establishes a semantic representation of the data
- system converts the plain English entry into SPARQL, queries the database, and displays the retrieved result

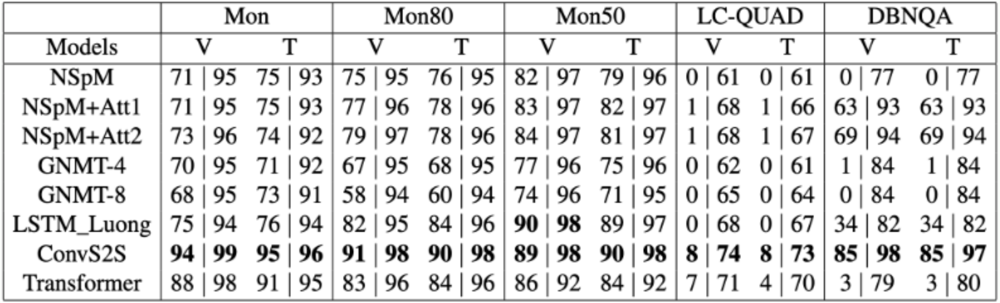
TNTspa

- Machine Translating from Natural Language to SPARQL.
- evaluating the utilization of eight different Neural Machine Translation(NMT) models
- the results show a dominance of a CNN-based architecture

Datasets:

- Monument
- Monument80
- Monument50
- LC-QUAD
- DBNQA

TNTspa



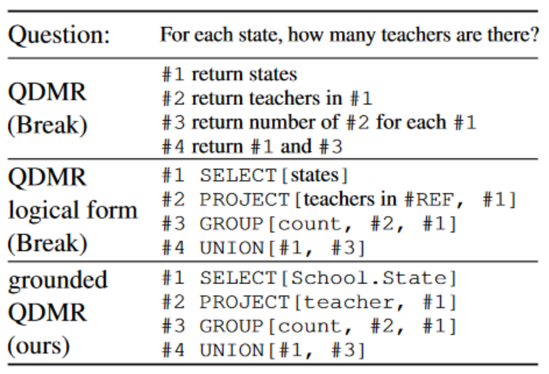
	Mon		Mon80		Mon50		LC-QUAD		DBNQA	
Models	V	T	V	T	V	T	V	T	V	T
NSpM	71 | 95	75 | 93	75 | 95	76 | 95	82 | 97	79 | 96	0 | 61	0 | 61	0 | 77	0 | 77
NSpM+Att1	71 | 95	75 | 93	77 | 96	78 | 96	83 | 97	82 | 97	1 | 68	1 | 66	63 | 93	63 | 93
NSpM+Att2	73 | 96	74 | 92	79 | 97	78 | 96	84 | 97	81 | 97	1 | 68	1 | 67	69 | 94	69 | 94
GNMT-4	70 | 95	71 | 92	67 | 95	68 | 95	77 | 96	75 | 96	0 | 62	0 | 61	1 | 84	1 | 84
GNMT-8	68 | 95	73 | 91	58 | 94	60 | 94	74 | 96	71 | 95	0 | 65	0 | 64	0 | 84	0 | 84
LSTM_Luong	75 | 94	76 | 94	82 | 95	84 | 96	**90** | **98**	89 | 97	0 | 68	0 | 67	34 | 82	34 | 82
ConvS2S	**94** | **99**	**95** | **96**	**91** | **98**	**90** | **98**	**89** | **98**	**90** | **98**	**8** | **74**	**8** | **73**	**85** | **98**	**85** | **97**
Transformer	88 | 98	91 | 95	83 | 96	84 | 96	86 | 92	84 | 92	7 | 71	4 | 70	3 | 79	3 | 80

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

Question Decomposition Meaning Representation

Natural Language questions into Logical

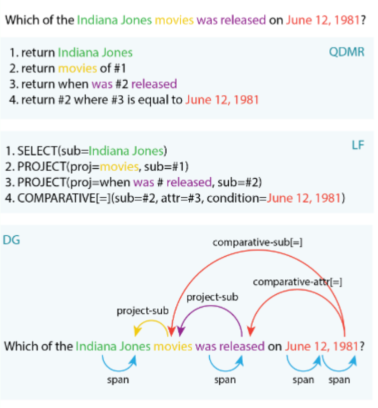
Intermediate representation for Natural Language questions.



Question:	For each state, how many teachers are there?
QDMR (Break)	#1 return states #2 return teachers in #1 #3 return number of #2 for each #1 #4 return #1 and #3
QDMR logical form (Break)	#1 SELECT[states] #2 PROJECT[teachers in #REF, #1] #3 GROUP[count, #2, #1] #4 UNION[#1, #3]
grounded QDMR (ours)	#1 SELECT[School.State] #2 PROJECT[teacher, #1] #3 GROUP[count, #2, #1] #4 UNION[#1, #3]

Figure: Wolfson et al. (2020)

Natural Language questions into Logical



Which of the Indiana Jones movies was released on June 12, 1981?

QDMR

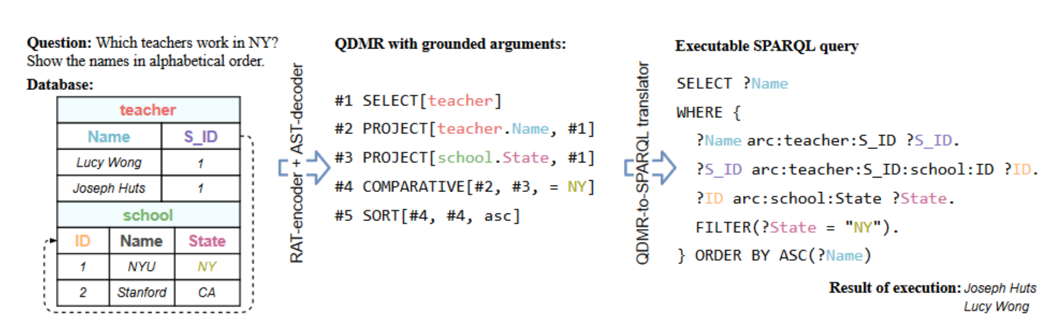
1. return Indiana Jones
2. return movies of #1
3. return when was #2 released
4. return #2 where #3 is equal to June 12, 1981

LF

1. SELECT(sub=Indiana Jones)
2. PROJECT(proj=movies, sub=#1)
3. PROJECT(proj=when was # released, sub=#2)
4. COMPARATIVE[=](sub=#2, attr=#3, condition=June 12, 1981)

Figure: Dependency Parsing of NL

QDMR to SPARQL



Question: Which teachers work in NY? Show the names in alphabetical order.

Database:

teacher	
Name	S_ID
Lucy Wong	*1*
Joseph Huts	*1*

school		
ID	Name	State
1	*NYU*	*NY*
2	*Stanford*	*CA*

QDMR with grounded arguments:

```
#1 SELECT[teacher]
#2 PROJECT[teacher.Name, #1]
#3 PROJECT[school.State, #1]
#4 COMPARATIVE[#2, #3, = NY]
#5 SORT[#4, #4, asc]
```

Executable SPARQL query

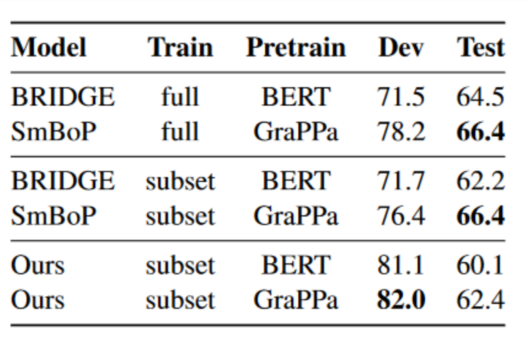
```
SELECT ?Name
WHERE {
  ?Name arc:teacher:S_ID ?S_ID.
  ?S_ID arc:teacher:S_ID:school:ID ?ID.
  ?ID arc:school:State ?State.
  FILTER(?State = "NY").
} ORDER BY ASC(?Name)
```

Result of execution: *Joseph Huts*
Lucy Wong

Figure: using QDMR to generate SPARCQL

QDMR to SPARQL

- Evaluation Metric: Execution Accuracy



Model	Train	Pretrain	Dev	Test
BRIDGE	full	BERT	71.5	64.5
SmBoP	full	GraPPa	78.2	**66.4**
BRIDGE	subset	BERT	71.7	62.2
SmBoP	subset	GraPPa	76.4	**66.4**
Ours	subset	BERT	81.1	60.1
Ours	subset	GraPPa	**82.0**	62.4

Figure: using QDMR to generate SPARCQL

Processing SPARCQL for execution

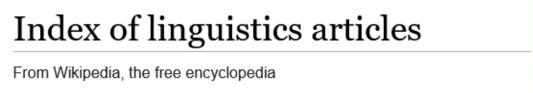
Processing Generated PARCQL

- Prefix Resolution
- Syntax issues
- Parenthesis

Resources

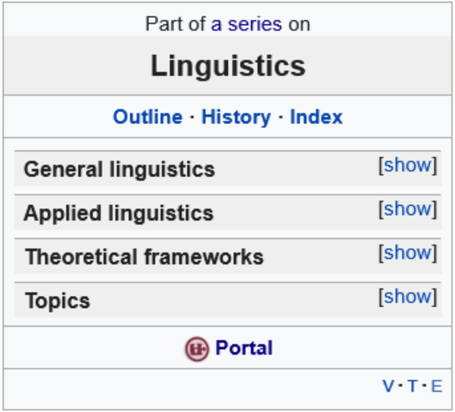
What is a part of linguistics?

For the purpose of this project: anything in ”Linguistics” category on Wikipedia.



Index of linguistics articles

From Wikipedia, the free encyclopedia



Part of a series on

Linguistics

Outline · History · Index

General linguistics [show]

Applied linguistics [show]

Theoretical frameworks [show]

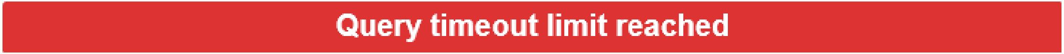
Topics [show]

Portal

V · T · E

Concepts

```
x is an instance of / subclass of* something studied by
linguistics
```



Query timeout limit reached

Decisions to make:

- include all instances of languages (dialects, jargon...)
- (at first) focus only on basic concepts (listed in Outline of linguistics)
- include parts of articles (definitions, examples) to enhance the knowledge graph

Thank you!

Any questions?