Master's Thesis

Answer Aggregation and Verbalization for Complex Question Answering

Presented by Zhiruo Zhang

Supervised by Ruijie Wang, Prof. Abraham Bernstein, Ph.D.

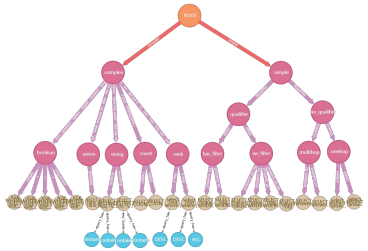
15.11.2023

1. Introduction

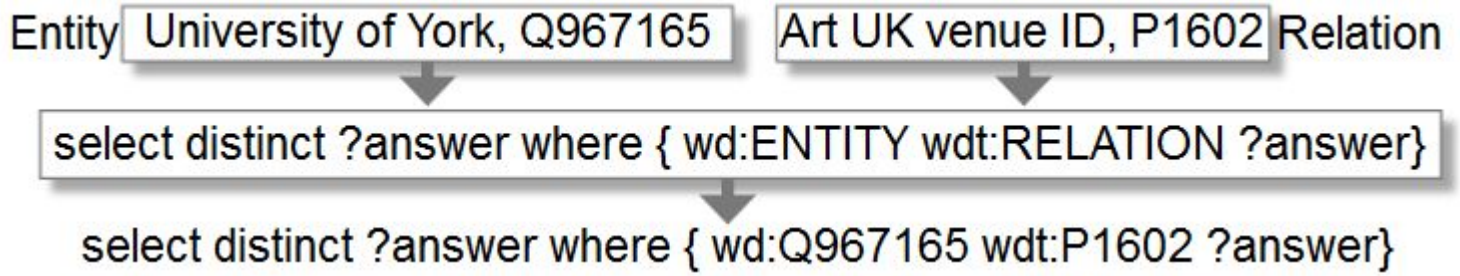
Background: complex question answering over knowledge graphs

Two Tasks:

- Question-to-SPARQL
  - Taxonomy tree construction



  - Query construction method (template-based)



  - Query generation method (BART-based)

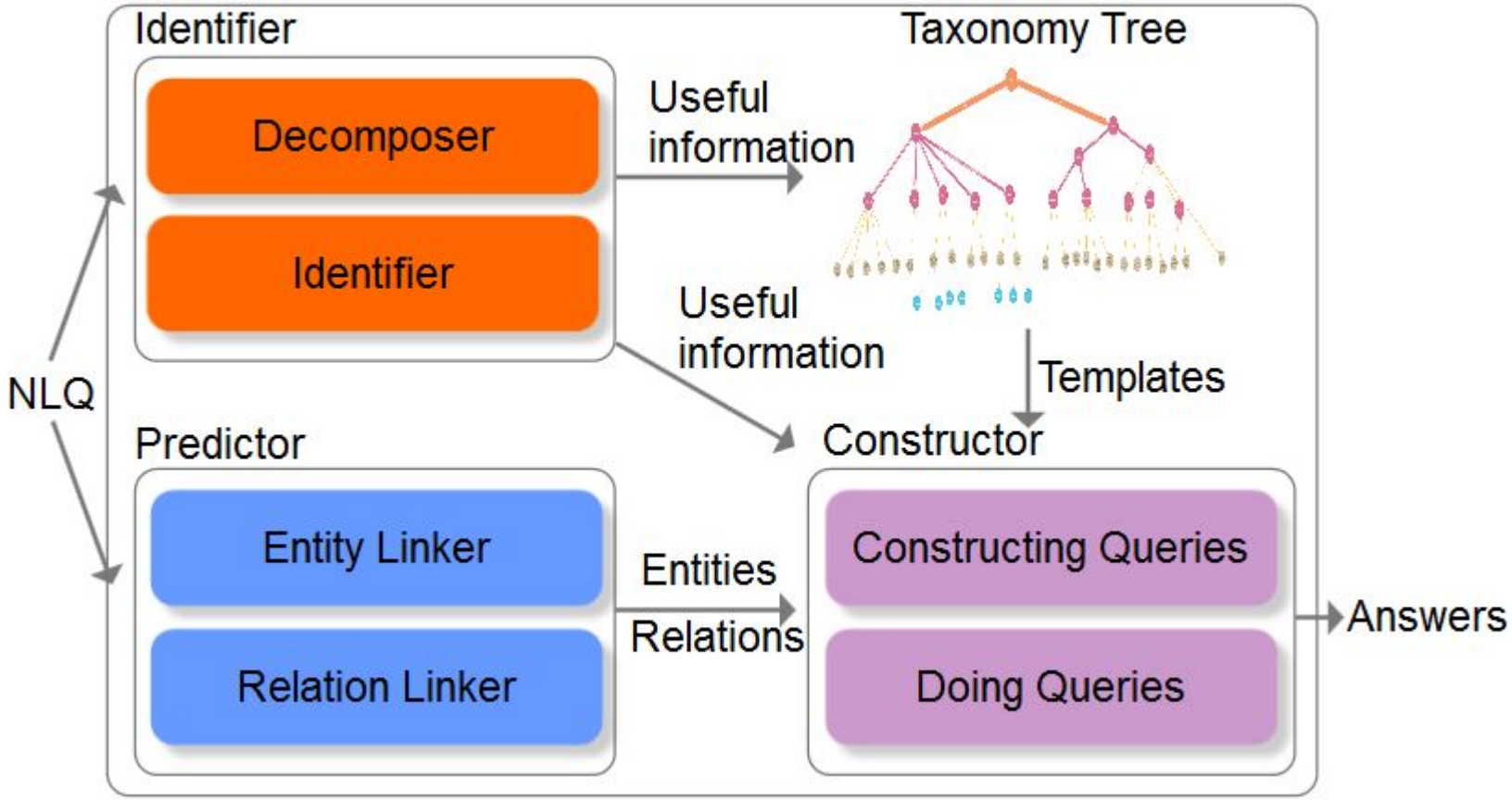
- Answer verbalization

  Question: What can be considered as category for The Spoiler ?
  Direct answer : comedy
  Verbalized answer: The Spoiler is considered as comedy category.

2. Methodology - Framework: Template-based

The template-based approach consists mainly of three modules:

- Identifier
  extracting useful information
- Predictor
  entity linking, relation linking
- Constructor
  constructing and querying



2. Methodology - Entity Linking

Question	Ground-truth Entities	DP	GENRE	SP
When was the GDP of Rio Grande do Sul1.15e+11 ?	Q40030: rio grande do sul	NULL	Q19879968: gdp, Q40030: rio grande do sul	Q314080: gewerkschaft der polizei, Q160636: rio grande
What is the name of the opera based on Twelfth Night ?	Q221221: twelfth night, Q1344: opera	Q221211: twelfth night	Q1344: opera	Q1071027: personal name, Q1344: opera Q221211: twelfth night
What feature is named after Juan de Fuca Plate, who died in the early 1600's?	Q860385: juan de fuca plate	Q860385: juan de fuca plate	NULL	Q42139305: feature, Q3391846:plate armor, Q107900: santa maria

Match

Existing methods:

- Spacy Entity Linker (SP)
  rule-based
- Entity Detection Module From DeepPavlov (DP)
  trained using BERT on LC-QuAD 2.0
- GENRE
  fine-tuned using BART on more than 20 datasets

Ranking entities by:

- Voting
- DP > GENRE > SP

2. Methodology - Relation Linking

Question	Ground-truth Relations	GenRL
When was the GDP of Rio Grande do Sul1.15e+11 ?	P2131: nominal GDP, P585: point in time	P585: point in time, P2131: nominal GDP, P176: manufacturer, P131: located in the administrative territorial entity, P4330: contains, P2079: fabrication method, P36: capital, P998: Curlie ID
What is the name of the opera based on Twelfth Night ?	P144: based on, P31: instance of	P144: based on, P87: librettist, P1877: after a work by, P86: composer, P136: genre, P1477: birth name, P676: lyrics by, P569: date of birth, P175: performer, P674: characters, P61: discoverer or inventor, P1433: published in, P179: part of the series, P570: date of death

Match

Redundancy

Not in the predictions

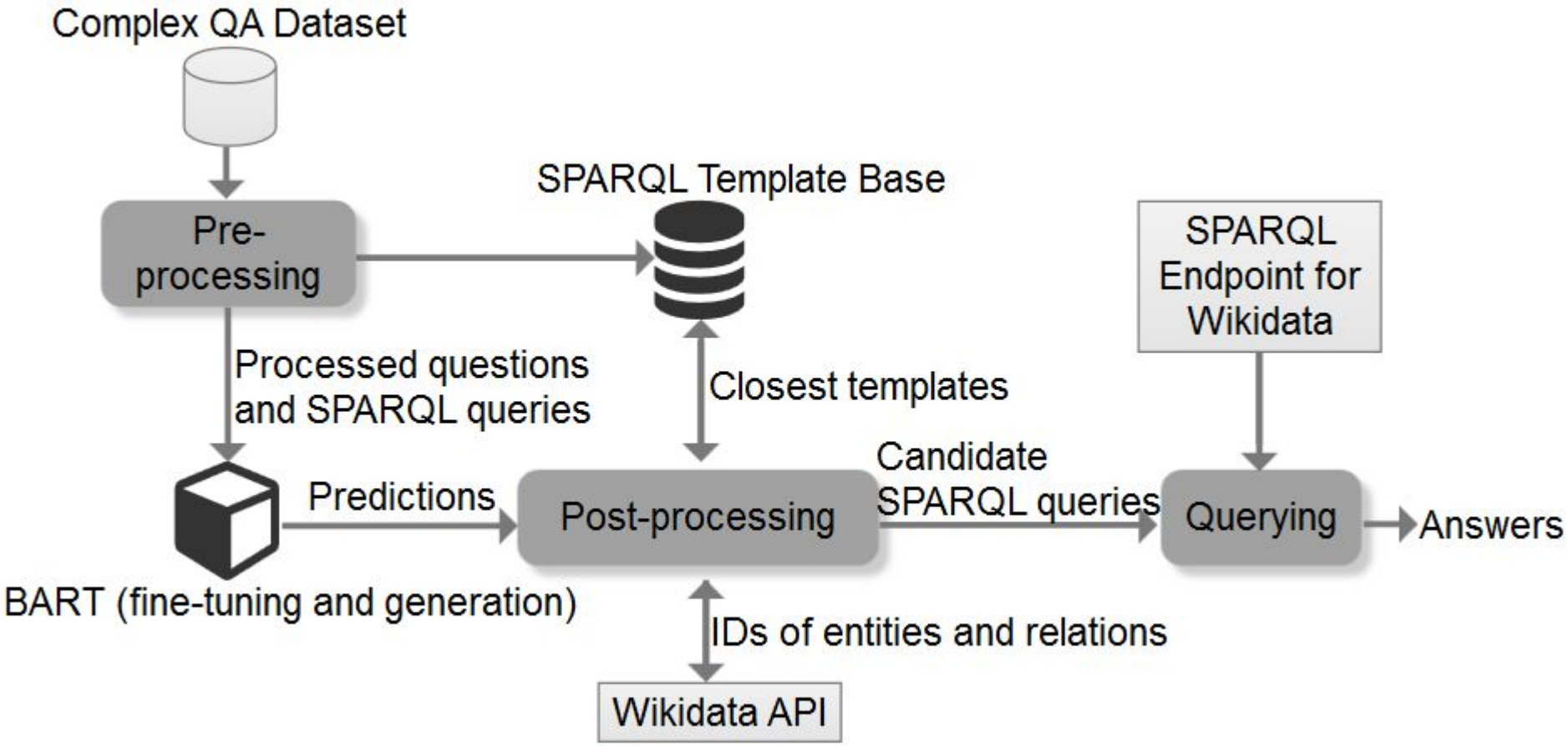
GenRL

- Fine-tuning BART
- Datasets
  LC-QuAD 1.0, LC-QuAD 2.0, QALD-9 and SimpleQuestions-WD

- Many relations are predicted
  - but often too many
  - missing can also happen

2. Methodology - Framework: BART-based

The BART-based approach consists of four modules:
Pre-processing, model fine-tuning/generation, post-processing and querying



2. Methodology - Pre-processing Queries

Ground–truth Query:

SELECT (COUNT(?sub) AS ?value) { ?sub wdt:P2546 wd:Q2695156 }

Query Template:

select ?count_of_sub bracket_open ?sub wdt: RELATION wd: ENTITY bracket_close

Processed Query:

select ?count_of_sub bracket_open ?sub wdt: sidekick_of wd: batman bracket_close

- Special symbols are replaced with special strings
- Some elements (e.g., count, filter, etc.) are simplified
- Entity IDs and relation IDs are replaced using
  - corresponding labels for its processed version
  - ENTITY, RELATION for its template

2. Methodology - Pre-processing Questions

Type	Input Question
original question	What is the size of the Andromeda Galax
ER_Within_Tag	what is the size of the wd: andromeda galax
ER_End_Tag	what is the size of the andromeda galax </s> wd: Andromeda </s> wd: elliptical_galaxy </s> wdt: child_astronomical_body </s> wdt: instance_of
ER_Within_End_Tag	what is the size of the wd: andromeda galax </s> wd: elliptical_galaxy </s> wdt: child_astronomical_body </s> wdt: instance_of

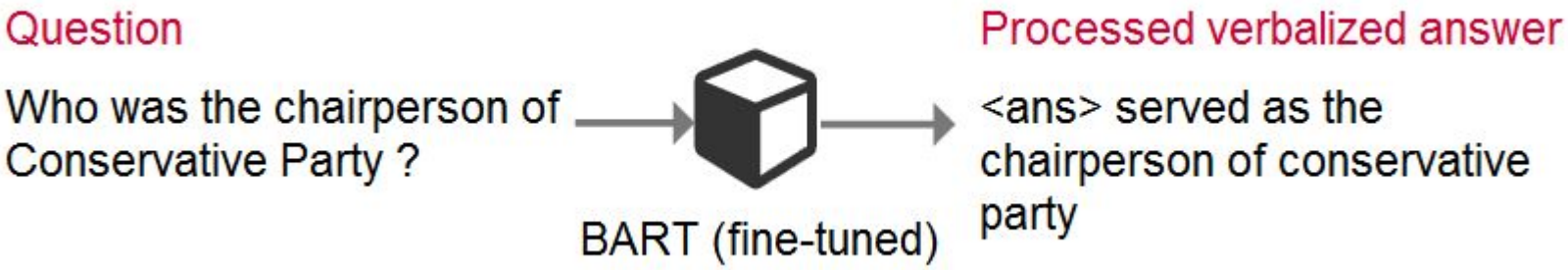
Mentioned in the question

Not explicitly mentioned

- Replacing some special symbols, such as parentheses
- Obtaining augmented inputs by attaching entity and relation labels to the question
  - ER_Within_Tag
  - ER_End_Tag
  - ER_Within_End_Tag

2. Methodology - Answer Verbalization

- Fine-tuning BART
- Questions are used as inputs for model fine-tuning
- Processed answers are used as the targets for model fine-tuning



2. Methodology - Evaluation Metrics

- Evaluation of SPARQL queries: Exact match (EM)
- Evaluation of answer verbalization:
  - BLEU
    A popular N-gram overlap metric to evaluate the similarity between two sentences
  - ROUGE-L
    - ROUGE calculates the recall score to evaluate the informativeness of generated response
    - ROUGE-L measures the longest common subsequence (LCS)
  - chrF

    chrF works on the granularity of character n-grams

3. Experiments - Datasets

LC-QuAD (Largescale Complex Question Answering Dataset) 2.0 was used for the question-to-SPARQL task

Wikidata as the KG

Dataset	Size of Training Set	Size of Test Set
LC-QuAD 2.0	24180	6046
Filtered LC-QuAD 2.0	20235	5125

The size of datasets in LC-QuAD 2.0

Dataset	#Entities in Training Set	#Entities in Test Set	Ratio of Entities in Intersection/in Test Set	#Relations in Training Set	#Relations in Test Set	Ratio of Relations in Intersection/in Test Set
LC-QuAD 2.0	19588	6691	0.5198	3171	1486	0.6878
Filtered LC-QuAD 2.0	17159	5773	0.4951	3047	1420	0.6768

Statistics of entities and relations in LC-QuAD 2.0

3. Experiments - Datasets

VQuAnDa and VANiLLA were used for answer verbalization generation

- VQuAnDa (Verbalization Question Answering Dataset) contains verbalized expressions of answers, constructed based on LC-QuAD 1.0
- VANiLLa (Verbalized Answers in Natural Language at Large scale) is mainly about simple questions.

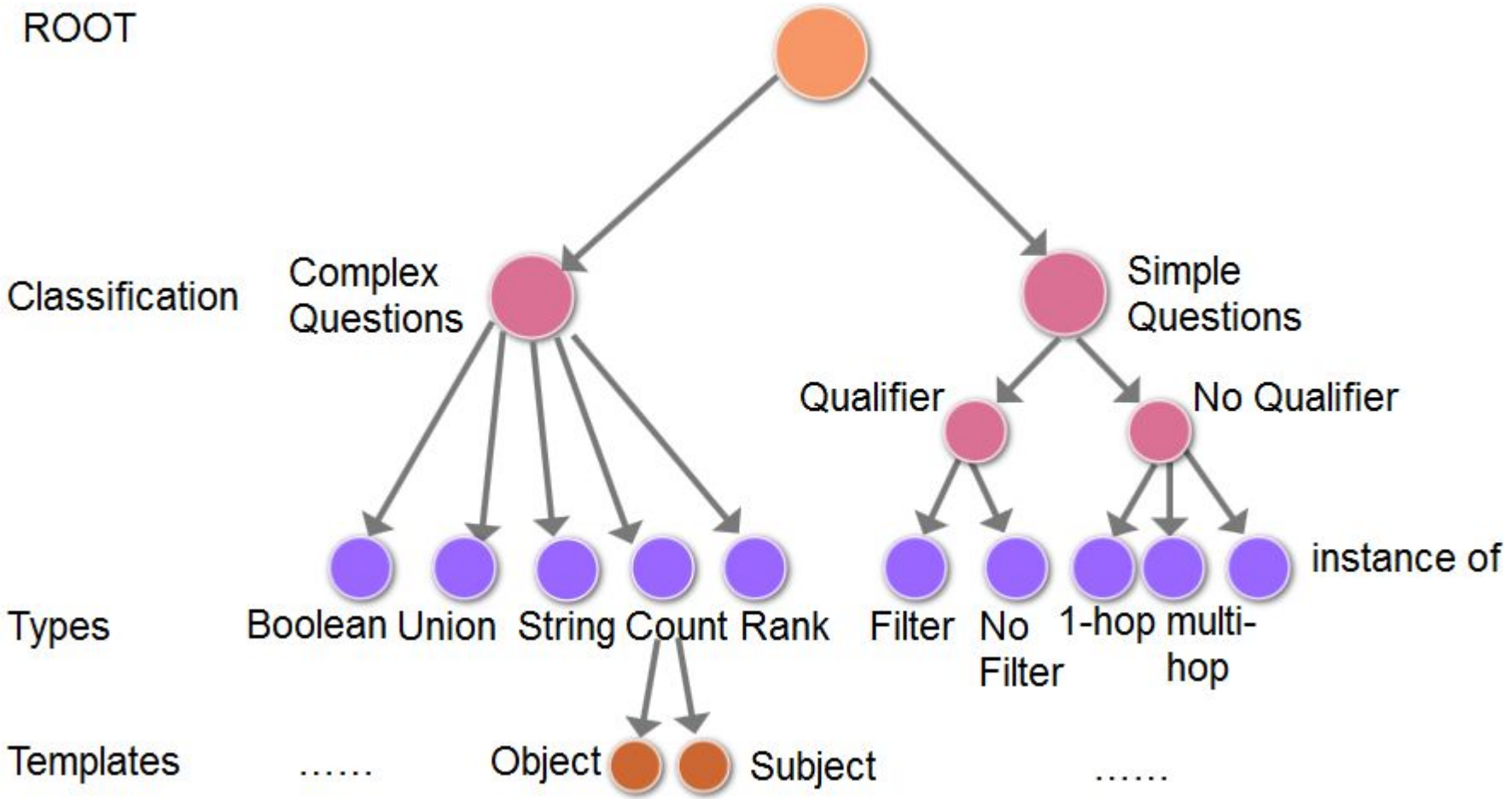
Dataset	Size of Training Set	Size of Test Set
VQuAnDa	4000	1000
VANiLLa	85732	21434

Statistics of VQuAnDa and VANiLLA

Dataset	Size of Training Set	Size of Validation Set	Size of Test Set
VANiLLa	85732	21434	10717
VANiLLa+VQuAnDa	89732	22434	11217

Statistics of datasets for fine-tuning

3. Experiments – Taxonomy Tree



- The node Qualifier means queries using p:, ps:, pq: to connect entities and relations, instead of the common wdt:

- Each Type node can be divided to several Template nodes. For example:
  - the templates of the count type include counting by subject or by object

3. Experiments – Results & Analysis

Performance of the template-based approach

Task	EM
entity linking	0.8256
relation linking	0.8033
entity linking & relation linking	0.6010
Query construction w/o searching	0.1947
Query construction w/ searching	0.1963

Poor performance due to:

- Limitations of existing entity linking and relation linking methods
- Misidentification of question types
  - typos
  - no typical words
- Difficulty to determine which template a simple question corresponds to

Question	Ground-truth Type
Name the painting that features Mona Lisa and that starts wit letter L	string
Who are the students of Pablo Picasso?	union
Is Paul Gascoigne a member of a sports team?	count

3. Experiments – Results & Analysis

Performance of the BART-based approach and comparisons

Method	Entities	Relations	Entities & Relations	Predicted Query
template-based w/ searching	0.8256	0.8033	0.6010	0.1963
question w/o tag	0.8264	0.7402	0.6534	0.5715
ER_Within_Tag	0.8549	0.7672	0.6931	0.5664
ER_End_Tag	0.9696	0.9824	0.9637	0.8046
ER_Within_End_tag	0.9571	0.9797	0.9509	0.7828
Diomedi and Hogan [2021]	-	-	-	0.1400
Zou et al. [2021]	-	-	-	0.5540

The overall performance of the BART-based approach is good and satisfying

- It has a 37% improvement over the template-based approach
- It performs better than previous works
- Augmented information is very helpful

3. Experiments – Results & Analysis

To validate the transferability of the BART-based approach to other KGs and QA datasets, additional experiments were done:

- DBLP-QuAD as the dataset
- DBLP scholarly knowledge graph as the KG

Dataset	Entities	Relations	Entities & Relations	Predicted Query
LC-QuAD 2.0	0.8264	0.7402	0.6534	0.5715
DBLP-QuAD	0.7100 ↓	0.8633 ↑	0.7087	0.6593

- The types of entities and their relations are stable in a scholar KG
- Entities, especially authors, are easily renamed and it is difficult to link to the expected one
- The statements are simpler, without the distinction between wdt:,p:,ps:,pq:

3. Experiments – Results & Analysis

Evaluation statistics of answer verbalization

Dataset	BLEU	ROUGE-L	chrF
VANiLLa	25.33	63.53	51.67
VANiLLa+VQuAnDa	24.02	64.83	55.03

The version VANiLLa+VQuAnDa

- has higher scores in chrF
- can give more appropriate answers to complex questions

An example count question and verbalized answers

Question:
How many people worshipped Yahweh?

Generation from Version VANiLLa: Bad
yahweh was worshipped by <ans>.

Generation from Version VANiLLa+VQuAnDa: Good
there are <ans> people who venerated yahweh.

4. Conclusions

The template-based approach

- Limited performance of existing entity linking and relation linking methods
- Errors in question identification and query construction
- Time-consuming, inefficient, low-accuracy

The BART-based approach

- Performed much better than the template-based approach
- Having trouble when dealing with unseen implicit entities and relations
  - useful augmentation information can help
- Good transferability

Answer verbalization

- Taking the context of the question into account
- Flexible and satisfying

Future works

Making the taxonomy tree and data open source

- Using other seq2seq auto-regressive LMs for fine-tuning and comparing their performances
- Applying the model-based method to other text-to-query tasks

The corresponding verbalized answer extensions of LC-QuAD 2.0

Thank you!