Question Answering on Scholarly Knowledge Graphs

Mohamad Yaser Jaradeh, Markus Stocker, Sören Auer

TPDL 2020 (Virtual), Full paper

Teaser!

- QA is widely researched topic
- Different techniques and technologies
- General domain knowledge
- No scholarly-oriented adoption
- No datasets, and no graphs

- Comes in JarvisQA

Introduction (1/2)

Hey Siri

alexa

"Ok Google"

yago
select knowledge

DBpedia

WIKIDATA

Freebase™

Introduction (2/2)

Why Scholarly knowledge is much more complicated to do QA on?

- Represented in unstructured manner
- Ambiguous
- Not FAIR
- Not machine actionable

Proposed Solution

Our proposed solution is **JarvisQA**

- BERT based system to answer questions on tabular views of scholarly knowledge graphs.
- Implemented on the ORKG[1] [1] scholarly knowledge graph

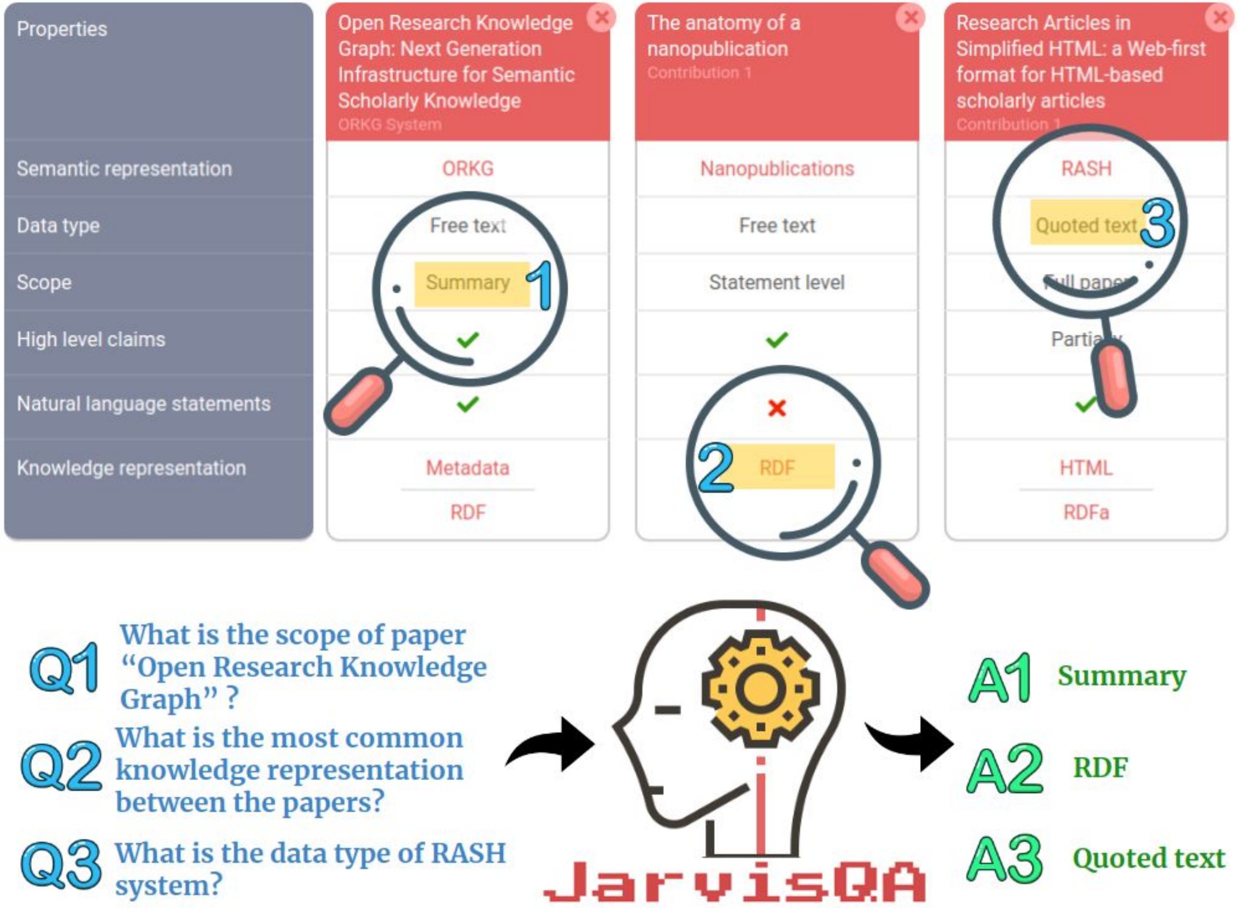
1. https://orkg.org

Related work

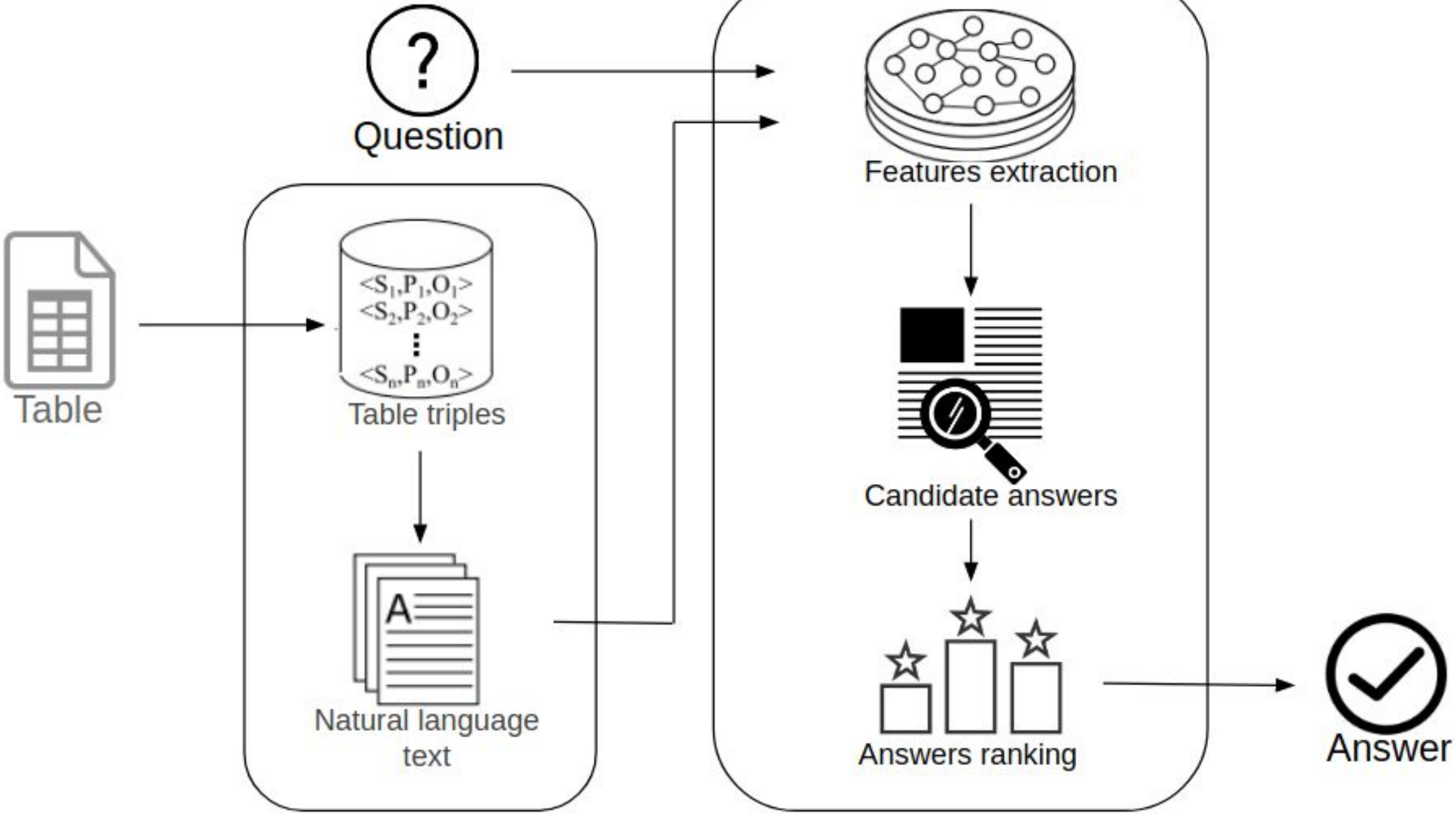
A plethora of work is done in the question answering domain.

- Frankenstein [2]
- QAnswer [3]
- ALBERT [4]
- Cheng et al. [5]
- TableQA [6]

JarvisQA


Properties
Semantic representation
Data type
Scope
High level claims
Natural language statements
Knowledge representation
Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge
ORKG System
ORKG
Free text
Summary
1
Metadata
RDF
The anatomy of a nanopublication
Contribution 1
Nanopublications
Free text
Statement level
2
RDF
Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles
RASH
Quoted text
3
HTML
RDFa
Q1
What is the scope of paper "Open Research Knowledge Graph"?
Q2
What is the most common knowledge representation between the papers?
Q3
What is the data type of RASH system?
JarvisQA
A1
Summary
A2
RDF
A3
Quoted text

How?


Question
Table
<S1,P1,O1>
<S2,P2,O2>
<Sn,Pn,On>
Table triples
Natural language text
Features extraction
Candidate answers
Answers ranking
Answer

Data (1/2)

We created the ORKG-QA dataset

- Compiled from data within the ORKG infrastructure
- Source of tables is ORKG comparisons
- 13 tables spanning 100+ publications
- 80 questions in English
  - Normal questions (one answer, one cell) **≈39%**
  - Aggregation questions (min, avg, most common, …) **≈20%**
  - Ask questions (True, false)
  - Listing questions (multiple results)
  - No answer questions (empty result)
  - **≈30%**
  - Complex questions (combining information) **≈11%**

Data (2/2)

For evaluation purposes, another dataset is used

- TabMCQ [7]
- “regents” tables
- Collected from 4th grader MCQ science exams
- 39 tables & 3745 questions

Evaluation (1/5)

Metrics used:

- Precision @k
- Recall @k
- F1-score @k
- Global Precision
- Global Recall
- Global F1-score
- Execution time
- Memory usage

Baselines used:

- Random
- Lucene[2]

2. https://lucene.apache.org/

Evaluation (2/5)

Experiment 1 (JarvisQA performance on the ORKG-QA dataset):

Questions type	Baseline	Precision @K				Recall @K				F1-Score @K			
		#1	#3	#5	#10	#1	#3	#5	#10	#1	#3	#5	#10
All	Random	0.02	0.06	0.08	0.16	0.02	0.07	0.09	0.18	0.02	0.06	0.08	0.17
All	Lucene	0.09	0.19	0.20	0.25	0.09	0.18	0.19	0.24	0.09	0.18	0.19	0.24
Normal	JarvisQA	0.41	0.47	0.55	0.61	0.41	0.47	0.53	0.61	0.41	0.47	0.54	0.61
Aggregation	JarvisQA	0.45	-	-	-	0.45	-	-	-	0.45	-	-	-
Related	JarvisQA	0.50	0.50	1.00	1.00	0.50	0.50	1.00	1.00	0.50	0.500	1.00	1.00
Similar	JarvisQA	0.11	0.25	0.67	-	0.11	0.25	0.67	-	0.11	0.25	0.67	-
All	JarvisQA	0.34	0.38	0.46	**0.47**	0.35	0.38	0.46	**0.48**	0.34	0.38	0.45	**0.47**

Evaluation (3/5)

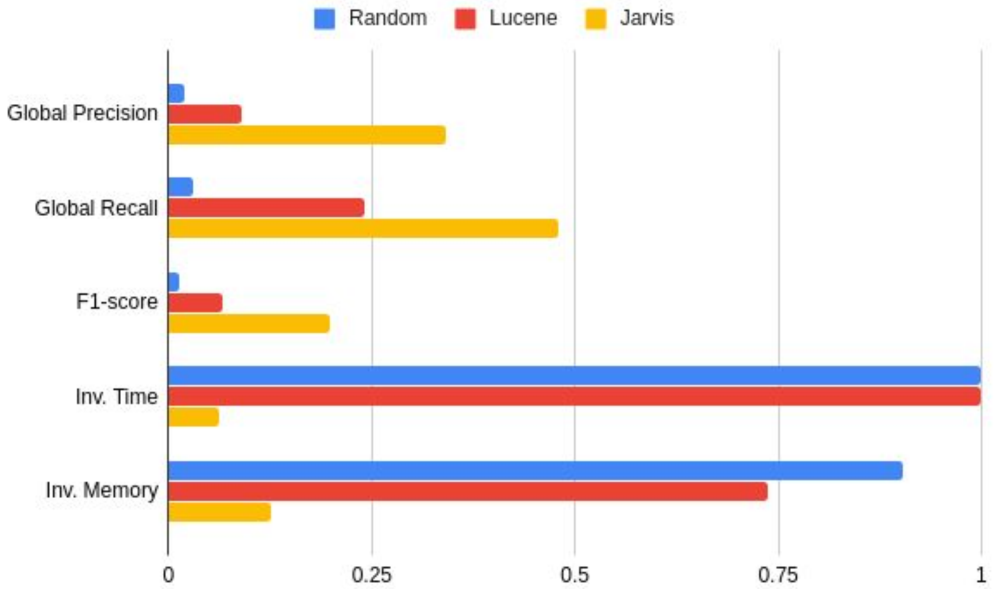
Experiment 2 (Different QA Models):

	Questions type	Precision @K				Recall @K				F1-Score @K			
		#1	#3	#5	#10	#1	#3	#5	#10	#1	#3	#5	#10
BERT L/U/S2	Normal	0.41	0.47	0.55	0.61	0.41	0.47	0.54	0.61	0.41	0.47	0.54	0.61
	Aggregation	0.45	-	-	-	0.45	-	-	-	0.45	-	-	-
	Related	0.50	0.50	**1.00**	-	0.50	0.50	**1.00**	-	0.50	0.50	**1.00**	-
	Similar	0.11	0.25	**0.67**	-	0.11	0.25	**0.67**	-	0.11	0.25	**0.67**	-
	All	0.35	0.38	0.46	0.48	0.35	0.38	0.46	0.48	0.34	0.38	0.46	0.48
ALBERT XL/S2	Normal	0.34	0.47	0.51	-	0.34	0.47	0.51	-	0.34	0.47	0.51	-
	Aggregation	0.45	0.45	**0.52**	-	0.45	0.45	**0.52**	-	0.45	0.45	**0.52**	-
	Related	**1.00**	-	-	-	**1.00**	-	-	-	**1.00**	-	-	-
	Similar	0.43	0.43	**0.67**	-	0.43	0.43	**0.67**	-	0.43	0.43	**0.67**	-
	All	0.36	0.42	0.46	-	0.37	0.43	0.47	-	0.36	0.42	0.46	-

B=Base; L=Large; XL=X-Large; C=Cased; U=Uncased; S1=Finetuned on SQuAD1; S2=Finetuned on SQuAD2

Evaluation (4/5)

Experiment 3 (metrics trade-off):



Evaluation (5/5)

Experiment 4 (performance on TabMCQ):

System	Dataset	Precision @K				Recall @K				F1-Score @K			
		#1	#3	#5	#10	#1	#3	#5	#10	#1	#3	#5	#10
Random	TabMCQ	0.006	0.010	0.020	0.030	0.010	0.020	0.030	0.040	0.007	0.010	0.024	0.030
	ORKG	0.020	0.060	0.080	0.160	0.020	0.070	0.090	0.180	0.020	0.060	0.080	0.017
Lucene	TabMCQ	0.004	0.018	0.027	0.036	0.006	0.017	0.026	0.037	0.005	0.016	0.024	0.033
	ORKG	0.090	0.190	0.200	0.250	0.090	0.180	0.190	0.240	0.090	0.180	0.190	0.240
Jarvis	TabMCQ	0.060	0.090	0.100	**0.110**	0.070	0.090	0.110	**0.120**	0.060	0.080	0.100	**0.110**
	ORKG	0.340	0.380	0.460	**0.470**	0.350	0.380	0.460	**0.480**	0.340	0.380	0.450	**0.470**

Discussion & Future Work

- Usual IR methods fail to find answers to questions
- JarvisQA outperforms other methods
- JarvisQA is a BERT-based system
  - Answers across multiple cells are an issue
  - True/False answers are an issue
  - Answers can be only as is in the text
- Future Work
  - Extend ORKG-QA dataset
  - Better answer selection
  - More question types support
  - Supplement tables with background knowledge

References

[1] M. Y. Jaradeh et al., “Open Research Knowledge Graph,” in Proceedings of the 10th International Conference on Knowledge Capture, 2019, doi: 10.1145/3360901.3364435.

[2] K. Singh et al., “Why Reinvent the Wheel,” in Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18, 2018, doi: 10.1145/3178876.3186023.

[3] D. Diefenbach et al., “QAnswer: A Question Answering prototype bridging the gap between a considerable part of the LOD cloud and end-users,” in The World Wide Web Conference on - WWW ’19, 2019, doi: 10.1145/3308558.3314124.

[4] Z. Lan et al., “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations” https://arxiv.org/abs/1909.11942

[5] J. Cheng et al., “Learning Structured Natural Language Representations for Semantic Parsing,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017.

[6] S. Vakulenko, V. Savenkov: TableQA: Question Answering on Tabular Data” (5 2017) https://arxiv.org/abs/1705.06504

[7] S.K. Jauhar et al. “TabMCQ: A Dataset of General Knowledge Tables and Multiple-choice Questions” (2 2016) https://arxiv.org/abs/1602.03960

Contact: <jaradeh@l3s.de>