LC-QuAD 2.0
Large-scale Complex Question Answering Dataset

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, Jens Lehmann
University of Bonn, Bonn, Germany
Fraunhofer IAIS, St. Augustin|Dresden, Germany
A goal properly set is halfway reached.
Outline

Motivation
LC-QuAD
Reified KGs
Objective
Process of Creating LC-QuAD 2.0
Outcome
Future Directions
Question Answering over KG

Question: “a sentence worded or expressed so as to elicit information.”

Definition: https://www.lexico.com/en/definition/question
Question Answering over KG

Question: “a sentence worded or expressed so as to elicit information.”

Definition: https://www.lexico.com/en/definition/question
Motivation
Motivation

a) Success of LC-QuAD 1.0

b) Impact of Reification of KG on KGQA
LC-QuAD
LC-QuAD

✔ has complex questions
✔ has SPARQL queries
✔ has boolean and aggregate based queries
✔ is supervised (gold standard)
✔ is extensible
✔ is awesome 😊
LC-QuAD

1885 Figshare views
45 Citations

QA Baselines
- QAmp [5]
- WDAqua [6]
- Krantikari QA [7]

QA Components
- EARL [8]
- Falcon [9]
- SQG [10]

Benchmarking:
- Hobbit Challenge [12]
- QALD-Gen [13]

Ensemble Systems
- Frankenstein QA [11]
Classical KG model
Reified KG Models
Reified KG Models

[Diagram showing nodes and edges with labels: subject, predicate, object, location, date.]

- Subject: spouse
- Predicate: location, date
- Object: 25 Aug 2005

[Image of characters: Robert Pattinson and Keira Knightley]
Outline

Motivation
LC-QuAD
Reified KGs

Objectives
Process of Creating LC-QuAD 2.0
Outcome
Future Directions
Objectives

To have a dataset with following property

**P1:** based on Reified KG model.

**P2:** high variety in complexity of questions.

**P3:** upscale the Complex QG process.

**P4:** to have paraphrased questions.
LC-QuAD
LC-QuAD 2.0
Dataset Creation Process
Workflow

1. Wikipedia Vital Entity list
   - SPARQL Template
   - Predicate list based on sparql
2. SPARQL
   - Template Fitting
   - Subgraph Generation
3. NNQT
   - Verbalisation
   - Papraphrasing
4. Final Verifiy

[Diagram]
Workflow
Workflow
Workflow
Workflow
Workflow
QT Template Question:
“What is {start time} and {place of marriage} of
{{Robert Downey Jr.} is {Spouse} of {Susan Downey}} ?”

25 Aug 2005
QT Template Question:
“What is {start time} and {place of marriage} of
{{Robert Downey Jr.} is {Spouse} of {Susan Downey}} ?”

QV Verbalised Question:
“When did Robert Downey Jr get married to
Susan Downey and at what place?”

25 Aug 2005
QT Template Question:
“What is {start time} and {place of marriage} of {{Robert Downey Jr.} is {Spouse} of {Susan Downey}} ?”

QV Verbalised Question:
“When did Robert Downey Jr get married to Susan Downey and at what place?”

QP Paraphrased Questions:
“When and where did RDJ got marry to Susan”
## Dataset(s)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Variation</th>
<th>Paraphrase</th>
<th>Target KG</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleQuestions [14]</td>
<td>100K</td>
<td>low</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>30M Factoid [15]</td>
<td>30M</td>
<td>low</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>QALD-9 [16]</td>
<td>450</td>
<td>high</td>
<td>No</td>
<td>DBpedia</td>
</tr>
<tr>
<td>Free917 [17]</td>
<td>917</td>
<td>medium</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>WebQuestions [18]</td>
<td>5K</td>
<td>medium</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>ComplexWebQuestionSP [19]</td>
<td>34K</td>
<td>medium</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>LC-QuAD 1.0 [1]</td>
<td>5K</td>
<td>medium</td>
<td>No</td>
<td>DBpedia</td>
</tr>
</tbody>
</table>
## Dataset(s)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Variation</th>
<th>Paraphrase</th>
<th>Target KG</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleQuestions [14]</td>
<td>100K</td>
<td>low</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>30M Factoid [15]</td>
<td>30M</td>
<td>low</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>QALD-9 [16]</td>
<td>450</td>
<td>high</td>
<td>No</td>
<td>DBpedia</td>
</tr>
<tr>
<td>Free917 [17]</td>
<td>917</td>
<td>medium</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>WebQuestions [18]</td>
<td>5K</td>
<td>medium</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>ComplexWebQuestionsSP [19]</td>
<td>34K</td>
<td>medium</td>
<td>No</td>
<td>Freebase</td>
</tr>
<tr>
<td>LC-QuAD 1.0 [1]</td>
<td>5K</td>
<td>medium</td>
<td>No</td>
<td>DBpedia</td>
</tr>
<tr>
<td><strong>LC-QuAD 2.0</strong></td>
<td><strong>30K</strong></td>
<td><strong>high</strong></td>
<td><strong>Yes</strong></td>
<td>Wikidata, DBpedia</td>
</tr>
</tbody>
</table>
LC-QuAD 2.0
Outline

Motivation
LC-QuAD
Reified KGs
Objectives
Process of Creating LC-QuAD 2.0

Outcome
Future Directions
Outline

Motivation
LC-QuAD
Reified KGs
Objectives
Process of Creating LC-QuAD 2.0

Outcome (did we achieve our objectives?)

Future Directions
Outcome
LCQuAD 2.0 is based on Reified KG model

SPARQL query following Reified KGs
Questions which ask for qualifier information
Questions with qualifier information
LCQuAD 2.0 have high variety in ques type

Classical complex questions (Multifacts)
Questions with dual intention
Questions with temporal aspects
Boolean, Count, Ranking
String operations
LCQuAD 2.0 have high variety in ques type

- Single Fact
  - Single Fact with type: 7.5%
  - Temporal: 10.2%
- Multi-fact: 23.9%
- Fact with qualifier: 10.2%
- Two intention: 7.8%
- Count: 4.8%
- Ranking: 8.9%
- String Operation: 8.9%
- Boolean: 4.1%
LCQuAD 2.0 upscale complex QG process

Transform the QG process to a AMT task

Template Question –> Verbalised Question
LCQuAD 2.0 have paraphrased questions

Able to have paraphrased questions in the dataset

QT $\rightarrow$ QV $\rightarrow$ QP

Thus, diverse vocabulary in the dataset
LCQuAD 2.0 have paraphrased questions
Objectives

To have a dataset with following property

- **P1**: based on Reified KG model.
- **P2**: high variety in complexity of questions.
- **P3**: upscale the Complex QG process.
- **P4**: to have paraphrased questions.
Discussion
Limitations

No UNION, OPTIONAL queries.

No out-of-scope questions.
Future Directions

KGQA Community to look into reified KGs.

Creating baselines.

Automatic question generation.
Acknowledgements

Fraunhofer-Cluster of Excellence (CCIT).

German Federal Ministry of Education and Research.
Acknowledgements

Dr. Michael Galkin
Senior Research Scientist
Enterprise Information System
Fraunhofer IAIS


16. 9th challenge on question answering over linked data (qald-9).
Questions?

Please feel free to AskNow
Website, Contacts

http://lc-quad.sda.tech

AskNowQA/LC-QuAD2.0
LC-QuAD 2.0
Largescale Complex Question Answering Dataset

What?

LC-QuAD 2.0 is a Large Question Answering dataset with 30,000 pairs of question and its corresponding SPARQL query. The target knowledge base is Wikidata and DBpedia, specifically the 2018 version. Please see our paper for details about the dataset creation process and framework.

Data Characteristics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Version</td>
<td>2.0</td>
</tr>
<tr>
<td>Total Questions</td>
<td>30,000</td>
</tr>
<tr>
<td>Unique Templates</td>
<td>22</td>
</tr>
<tr>
<td>Entities Covered</td>
<td>21,258</td>
</tr>
<tr>
<td>Predicates Covered</td>
<td>1,310</td>
</tr>
</tbody>
</table>

Leaderboard

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Exact Match</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
</tbody>
</table>

Usage