Benchmarking Entity Linking for Question Answering over Knowledge Graphs

Guillermo Echegoyen Blanco
Álvaro Rodrigo
Anselmo Peñas
{gblanco, alvarory, anselmo} **at** lsi.uned.es

NLP & IR Group
Universidad Nacional de Educación a Distancia

Overview

1. Introduction
2. Benchmark
3. Results
4. Outcomes
5. Conclusions & Future Work
6. References

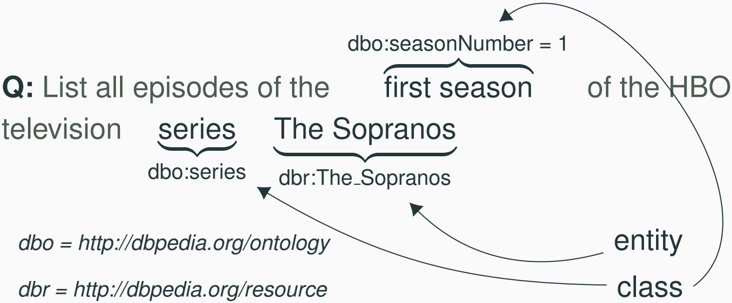
Introduction

Introduction

Entity Linking

Def: Link parts of a Natural Language passage to their corresponding node in a Knowledge Graph. Usually comprises:

- Recognize the entity mention in the text.
- Disambiguate the mention.



Q: List all episodes of the first season of the HBO television series The Sopranos

dbo = http://dbpedia.org/ontology

dbr = http://dbpedia.org/resource

Introduction

Motivation

- Lots of QA systems do perform an EL step with good results.
- Is the task easy or datasets are?
- Asses impact of EL Task on QA systems over KG.
- Actual collections for QA are easy for Entity Linking.

Q: List all episodes of the of the HBO television series

$$\underbrace{\text{The Sopranos}}_{\text{dbr:The\_Sopranos}}$$

replace " " → "_"

Benchmark

Benchmark

Objective: Complex dataset for Entity Linking

Input Datasets

- QALD $\{$1-4$\}$ Unger et al. 2014) $\leq$ 200 QA pairs each
- LC-QuAD (Trivedi et al. 2017) 5K QA pairs

Benchmark

Example

```
{
  "id": "37",
  "query": { "sparql": "SELECT ?uri ... },
  "answers": {
    "answer": [{ ...
    }, ...]
  },
  "question": [
    {
      "string": "List all episodes of the first season of the
          HBO television series The Sopranos!",
      "language": "en"
    }
  ]
}
```

Benchmark # Characterization

Difficulty?

- Given the Question, how easy is the Entity Linking?

Cases	**QALD-1**	**QALD-2**	**QALD-3**	**QALD-4**
Identical to DBP uri	**73.33%**	**85.71%**	**84.27%**	**79.21%**
Missing tokens		4.76%	5.62%	9.9%
Additional tokens	20.0%	1.19%	1.12%	0.5%
Lexical variation	6.67%	5.95%	5.62%	8.42%
Other		2.38%	3.37%	1.98%

Benchmark # Automatic Generation

Strategy

1. Develop method to detect easy mentions
2. Remove easy mentions from collection

Methods

- Trigram based mention detection
- Distance based mention detection

Results

Results

Developed method removed between 50% and 70% of each dataset.

Released Datasets

- **QALD-$\{$1-4$\}$-EL**: QALD-X version for EL $\leq$ 45 samples each.
- **LC-QuAD-EL**: LC-QuAD version for EL $\leq 1.3K$ samples.
- **C-EL4QA**: Complex compilation of EL versions $\leq 1.5K$ samples.

Outcomes

Outcomes

Our main contributions are:

- QA Datasets characterization
- Semi-automatic method to generate EL dataset.
- Release large benchmark dataset and baseline for EL in QA.

Conclusions & Future Work

Conclusions & Future Work

Conclusions

- We found QA collections to be very easy
- QA Systems go for automated solutions

Research Questions

- If Entity Linking were more difficult, how QA system would perform?
- How can we create more difficult Entity Linking collections?

Thank you!

Questions?

References

References

Priyansh Trivedi et al. “Lc-quad: A corpus for complex question answering over knowledge graphs”. In: *International Semantic Web Conference*. Springer. 2017, pp. 210–218.

Christina Unger et al. “Question Answering over Linked Data (QALD-4)”. In: (2014). URL: `https://hal.inria.fr/hal-01086472/`.