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. Characterization
3. Automatic Generation
4. Results
5. Conclusions & Future Work
6. Outcomes
7. References

Introduction

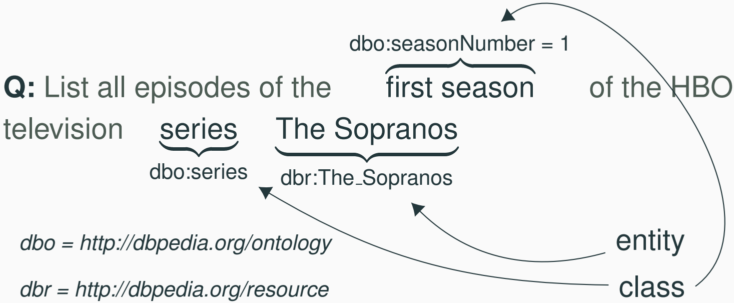
Introduction

| UNED

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

Motivation

- Lots of QA systems do perform an EL step with good results.
- Asses impact of EL Task on QA systems over KG.

Characterization

Characterization

Input Datasets

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

Characterization

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"
    }
  ]
}
```

Characterization

Difficulty?

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

Cases	**QALD-1**	**QALD-2**	**QALD-3**	**QALD-4**
Identical to DBP uri	**92.0%**	**72.0%**	**75.0%**	**80.0%**
Missing tokens		4.0%	5.0%	10.0%
Additional tokens	6.0%	1.0%	1.0%	0.5%
Lexical variation	2.0%	5.0%	5.0%	8.5%
Other		18.0%	14.0%	1.0%
Distance method	92.0%	80.0%	83.0%	89.5%
Trigram method	92.0%	84.0%	86.0%	94.5%

Characterization

Problem

- 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 ” ” → ”_”

Automatic Generation

Automatic Generation

Objective: Complex dataset for Entity Linking

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

Released Datasets

Dataset	Unique Q.	Unique E.	Total
QALD-1-EL	3	3	4
QALD-2-EL	11	11	12
QALD-3-EL	13	13	14
QALD-4-EL	38	40	45
LC-QuAD-EL	1204	997	1292
C-EL4QA	1269	1064	1367

Conclusions & Future Work

Conclusions & Future Work

Conclusions

- We found QA that collections do not really tackle the EL problem.
- QA Systems go for automated solutions

Open Questions

- If Entity Linking were more difficult, how QA system would perform?

Outcomes

Outcomes

Our main contributions are:

- QA Datasets characterization
- Semi-automatic method to generate complex EL datasets.
- Release large benchmark dataset and baseline for EL in QA (url)

Thank you!

Questions?

References

References i

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/`.