Core techniques of QA Systems over KBs a Survey

Guillermo Echegoyen Blanco

Summary

Here goes the Summary

Intro

- A Question Answering System should be able to: *Understand a Natural Language Question so as to be able to answer based on some pre-known data.*

- Typically involves accepting a question and generating a SparQL query capable of extracting the information which answers the user question.

- QALD benchmark
- WebQuestions benchmark
- SimpleQuestions benchmark

Tasks

- Question Analysis
- Phrase Mapping
- Disambiguation
- Query Construction
- Distributed Knowledge

Question Analysis #1

Analyze syntactic features to extract meaningful information:

- Type of question (is it a Which, What. . . question).
- Multilinguality (is it in English, French. . .).
- Correspondance to KB entities/classes.
- Tokens in the sentence and it's relations.
- Useless words in the sentence.

Question Analysis #2

Techniques based on:

- Recognizing Named Entities
- Segmenting with *POS** Tags
- Identifying dependencies using parsers

POS Tag: Part-Of-Speech Tag

Question Analysis #3 - Recognizing named entities

Identify Named Entities and map to resource in KB

- *NER* Tools: Tools from NLP, **Standford NER Tool**. Domain specific, **low precision 51%** (He et al. 2014)
- *N-Gram*: Map n-grams to KB entities. Adv: Each NE can be recognized in the KB, disadv: Dissambiguation explodes (**too much candidates**). (SINA: Shekarpour et al. 2015, CASIA: He et al. 2014)
- *Entity Linking* Tools: **DBpedia Spotlight** (Daiber et al. 2013) and **AIDA** (Yosef et al. 2011). Recognize NE and find the underlying KB resource, dissambiguating on the way. Adv: All-in-one. Disadv: Limited service, **KB dependant**.

Question Analysis #4 - Segmenting using POS Tagging

Identify which phrase correspond to instances, properties, classes. . . and which is irrelevant.

- *Handmade rules*: Regular expressions depenending on question type, structure. . . . (PowerAqua Lopez et al. 2012, Treo Freitas and Curry 2014, DEANNA Yahya et al. 2013). Disadv: **regex built by hand**.
- *Learning rules*: **Machine Learning** approach, train over corpus (Xser Xu, Feng, and Zhao 2014, UTQA "Pouran-ebn veyseh A" 2016). Disadv: **training corpus needed**.

Question Analysis #5 - Parsers

Grammar based parsers to generate trees or DAGs

- *Dependency grammars*: **Standford dependency parser**, word dependencies. Adv: can extract relations along with it's arguments (gAnswer Zou et al. 2014, **PATTY** Nakashole, Weikum, and Suchanek 2012)
- *Dependencies and DAGs*: Dependencies between phrases. Disadv: **parser trained on dataset** (Xser Xu, Feng, and Zhao 2014).

Question Analysis #6 - Summary

Which techniques to choose?

- Xser (**trained DAG**) reports best results on *QALD 4.1 & 5*
- gAnswer (**Dependency grammars**) reports fastest results on *QALD 3 & 4*

Machine Learning approach: Can be fast enough and there is plenty of data available.

Phrase Mapping #1

Find the resources in the KB with the highest probability that maps to the phrase.

Problems:

- String similarity
- Semantic similarity
- Language

Phrase Mapping #2

- Database with lexicalization: *WordNet, Wiktionary, PATTY* Expand the phrase with synonims and use that for search. Adv: High number of candidates, disadv: **Big search space**, **not very useful for domain specific mappings**.
- Mappings using large texts: **word2vec** semantics reflected in the associated vector. Adv: aids in the **lexical gap**, disadv: **needs training on large texts, noisy, performance**.

Disambiguation

Query Construction

Distributed Knowledge

Bibliography

“Pouran-ebn veyseh A”. fr. In: In: ESWC. to appear. 2016.

J. Daiber et al. “Improving efficiency and accuracy in multilingual entity extraction”. en. In: *Proceedings of the 9th international conference on semantic systems, ACM*. 2013.

A. Freitas and E. Curry. “Natural language queries over heterogeneous linked data graphs: a distributional-compositional semantics approach”. en. In: *Proceedings of the 19th international conference on intelligent user interfaces, ACM*. 2014.

S. He et al. “CASIA@ V2: a MLN-based question answering system over linked data”. nl. In: *Proceedings of QALD-4*. 2014.

V. Lopez et al. “Poweraqua: supporting users in querying and exploring the semantic web”. en. In: