Core techniques of

QA Systems over KBs a Survey

Guillermo Echegoyen Blanco

Overview

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), **DBpedia Lookup** 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 **lexical gap, string similarity and semantic similarity**, disadv: **needs training on large texts, noisy, performance**.

Phrase Mapping #3 - Summary

Which techniques to choose?

ToDo

Disambiguation #1

QA systems generate lots of possible interpretations due to language ambiguities and search process.

- Find univocally the resource that maps to the requested question.

Typically approached:

- String or semantic similarity to resource label (include).
- Consistency check between the properties and their arguments (exclude).

Disambiguation #2 - Graph Search

Dissambiguation carried out in the KB search step

- Subgraph matching against the KB (**gAnswer** Zou et al. 2014 does it on phrase mapping). Represent the question as a dependency graph and find an isomorfic subgraph in KB. Adv: very fast. Disadv: dissambiguation carries over. (**high precision, low recall**)
- Search both with edges and nodes (**PowerAqua** Lopez et al. 2012). Disadv: slow.
- **SemSek** Aggarwal and Buitelaar 2012 and **Treo** Freitas and Curry 2014 do it only with recognized instances. (**low precision, high recall**)

Dissambiguation #3 - Graph Search

gAnswer searches in the edges and vertices, **SemSek, Treo** search on instances and properties attached.

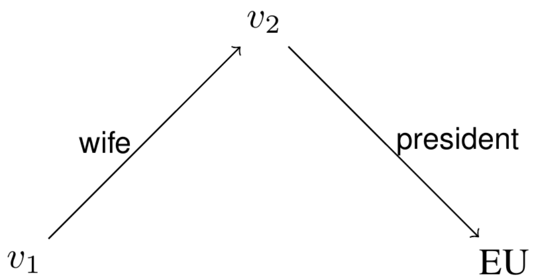


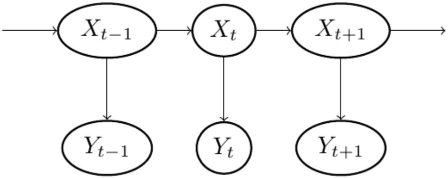
Figure: Subgraph generated for the question "Who is the wife of the president of the EU?"

Dissambiguation #4 - Hidden Markov Model (HMM)

"By which countries was the EU founded?"

Two stochastic processes:

- **Hidden (dissambiguation)** $X_{t \in N}$**:** $\{$*dbo:Country, dbr:Euro, dbr:European_Union, dbp:founded, dbp:establishedEvent*$\}$.
- **Observed (question tokens),** $Y_{t \in N}$**:** $\{$*"countries", "EU", "founded"*$\}$.



Dissambiguation #5 - HMM

The problem is reduced to find the most probable set of states. Extra parameters:

- Initial probability $P(X_0 = x)$ for $x \in X$
- Transition probability $P(X_t = x_1 | X_{t-1} = x_2)$ for $x_1, x_2 \in X$
- Emission probability $P(Y_t = y | X_t = x)$ for $x \in X, y \in Y$

It is not necessary to know the the dependency between different resources, just the available resources.

Disambiguation #6 - HMM

SINA (Shekarpour et al. 2015): **slow**

- Emission: string similarity between label and segment.
- Initial & Transition: estimated based on the distance of the resource in the KB and popularity.

RTV (Giannone, Bellomaria, and Basili 2013): **inaccurate**

- Emission: word embeddings
- Initial & Transition: uniform across all resources

Dissambiguation #7 - ILP & MLN

ILP Optimization problem

- **DEANNA** (Yahya et al. 2013) Dependencies between the segments have to be computed in the question analysis phase. **slow, low precision & recall**

Markov Logic Network

- **CASIA** (He et al. 2014) Hard constraints like ILP, soft constraints flexibility **training needed low precision & recall**

Dissambiguation #8 - Structured Perceptron

Considering:

- Similarity of the phrase and the corresponding resource
- Popularity of a label for a resource
- Compatibility of the range and domain of a property with the arguments.

Xser (Xu, Feng, and Zhao 2014) Solves ambiguity **fast, training needed**

Dissambiguation #9 - Summary

Which techniques to choose

ToDo

Query Construction

Distributed Knowledge

References I

- "Pouran-ebn veyseh A". fr. In: In: ESWC. to appear. 2016.
- N. Aggarwal and P. Buitelaar. "A system description of natural language query over dbpedia. In: Proceedings of interacting with linked data". en. In: ILD, 2012.
- 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.

References II

C. Giannone, V. Bellomaria, and R. Basili. "A HMM-based approach to question answering against linked data. In: Proceedings of the question answering over linked data". en. In: lab (QALD-3) at CLEF. 2013.

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: *Semant Web* 3.249–265 (2012).

N. Nakashole, G. Weikum, and F. Suchanek. "PATTY: a taxonomy of relational patterns with semantic types". en. In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, association for computational linguistics*. 2012.

S. Shekarpour et al. *Sina: semantic interpretation of user queries for question answering on interlinked data. Web Semant Sci Serv Agents World Wide Web 30(Supplement C):39–51*. en. 2015. DOI: doi:10.1016/j.websem.2014.06.002.

K. Xu, Y. Feng, and D. Zhao. "Xser@ QALD-4: answering natural language questions via phrasal semantic parsing". pt. In: *Natural Language Processing and Chinese Computing*. Springer, 2014, pp. 333–344.

References IV

M. Yahya et al. “Robust question answering over the web of linked data”. en. In: *Proceedings of the 22nd ACM international conference on conference on information knowledge management, ACM*. 2013.

M.A. Yosef et al. “Aida: An online tool for accurate disambiguation of named entities in text and tables”. en. In: *Proceedings of the VLDB*. 2011.

L. Zou et al. “Natural language question answering over RDF: a graph data driven approach”. en. In: *Proceedings of the 2014 ACM SIGMOD international conference on Management of data, ACM*. 2014.