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

Question Analysis #1

Analyze syntactic features to extract meaningful information:

- Type of question
- Multilinguality
- 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 dependending on question type, structure. . . . (PowerAqua Lopez et al. 2012, Treo Freitas and Curry 2014, DEANNA Yahya et al. 2013). Disadv: **regex built by hand**.

WRB	VBD	DT	NNP	NNP	VBN	.
When	was	the	European	Union	founded	?

Figure: POS tagging from the **Standford POS Tagger**

Question Analysis #5 - Segmenting using POS Tagging

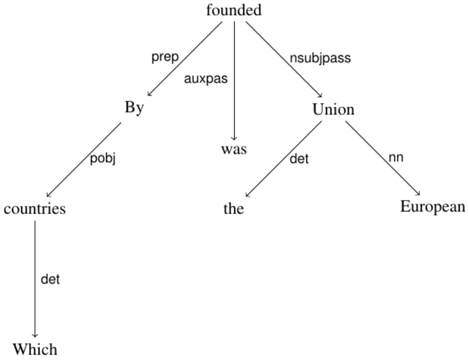
- *Learning rules*: **Machine Learning** approach, train over corpus (Xser Xu, Feng, and Zhao 2014, UTQA "Pouran-ebn veyseh A" 2016, very good results). Disadv: **training corpus needed**.

none	V-B	C-B	none	none	E-B	E-I	R-B	.
By	which	countries	was	the	European	Union	founded	?

Figure: Question annonated with **CoNLL IOB format**

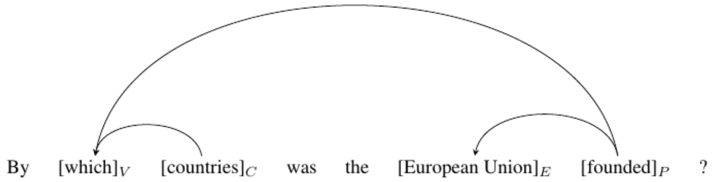
Question Analysis #6 - Parsers

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)



Question Analysis #7 - Parsers

Phrase Dependencies and DAGs: Dependencies between phrases. SHIFT-REDUCED parser. Disadv: **parser trained on dataset** (Xser Xu, Feng, and Zhao 2014).



Question Analysis #8 - Summary

	NER	NE n-gram startegy	EL tools	POS hand-made	POS learned	Parser structural grammar	Dependency parser	Phrase dependencies and DAG
BELA				x				
CASIA		x			x		x	
DEANNA		x		x			x	
FREyA						x		
gAnswer							x	
GFMed								
Hakimov et al.								x
Intui2						x		
Intui3	x			x				
ISOFT			x	x			x	
POMELO		x						
PowerAqua				x				
QAKiS	x	x						
QAnswer		x					x	
RTV	?						x	
SemGraphQA		x					x	
SemSeK		x				x	x	
SINA		x						
SWIP	?						x	
TBSL				x				
TR Discover								
Treo				x			x	
UTQA					x			
Xser	?				x			x
Zhang et al.		x						
Zhu et al.						x		

Question Analysis #9 - 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*
- UTQA (**Learned POS tags**) reports best results on *QALD 6*

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

- KB Labels: Search in the labels provided by the KB's entitiy (all)
- Redirects: Follow the *owl:sameAs* links (gAnswer Zou et al. 2014)
- Extracted knowledge: From the previous phase (gAnswer Zou et al. 2014). Relations and arguments.
- Wikipedia specific: **DBPedia Lookup**, **Wikimedia Miner Tool** (gAnswer Zou et al. 2014, Xser Xu, Feng, and Zhao 2014, Zhu et al. n.d.)

Phrase Mapping #4 - Summary

	Knowledge base labels	String similarity	Lucene index or similar	WordNet/Wiktionary	Redirects	PATTY	Using extracted knowledge	BOA or similar	Distributional Semantics	Wikipedia specific approaches
BELA	x	x	x	x	x				x	
CASIA	x				x		x		x	
DEANNA	x									
FREyA	x	x		x						
gAnswer	x				x		x			x
GFMed	x				x					
Hakimov et al.	x							x		
Intui2	x									
Intui3	x			x	x					x
ISOFT	x		x			x			x	
POMELO	x									
PowerAqua	x		x	x	x					
QAKiS	x							x		
QAnswer	x	x	x	x	x			x		
RTV	x		x						x	
SemGraphQA	x		x	x	x					
SemSeK	x		x	x	x				x	
SINA	x									
SWIP	x	x								
TBSL	x	x	x	x				x		
TR Discover	x									
Treo	x		x						x	
UTQA	x	x			x				x	
Xser	x					x				x
Zhang et al.	x	x								
Zhu et al.	x								x	x

Phrase Mapping #5 - Summary

Which techniques to choose?

The best results here depend on the previous step and the computing resources available.
Options (can be combined together):

- KB's Labels
- Redirects
- word2vec
- PATTY

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.

Base approach (local dissambiguation):

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

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**)
- **SemSek** Aggarwal and Buitelaar 2012 and **Treo** Freitas and Curry 2014 do it only with recognized instances (during question analysis phase). (**low precision, high recall**)

Assume that all relational phrases can be deduced from the question.

Dissambiguation #3 - Graph Search

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

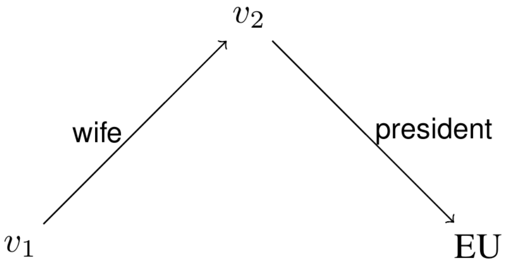


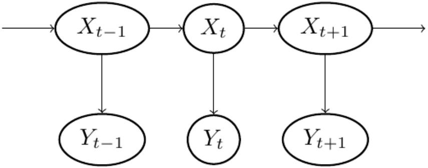
Figure: Subgraph generated for the question "Who is the wife of the president of the EU?"

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

	Local disambiguation	Graph search	HMM	LIP	MLN	Structured perceptron	User feedback
BELA	x						
CASIA	x				x		
DEANNA	x			x			
FREyA	x						x
gAnswer	x	x					
GFMed	x						
Hakimov et al.	x						
Intui2	x						
Intui3	x						
ISOFT	x						
POMELO	x						
PowerAqua	x	x					
QAKiS	x						
QAnswer	x						
RTV	x		x				
SemGraphQA	x						
SemSeK	x	x					
SINA	x		x				
SWIP	x						x
TBSL	x						
Treo	x	x					
TR Discover	x						
UTQA	?						
Xser	x					x	
Zhang et al.	x						
Zhu et al.	x						

Dissambiguation #9 - Summary

Which techniques to choose

- Best results *QALD 6*, reported from UTQA ("Pouran-ebn veyseh A" 2016), whose method is unkown.
- Fastest method is gAnser (Zou et al. 2014) which dissambiguates in the phrase mapping step. (**subgraph matching**)
- Best results in *QALD 4.1 & 5* by Xser (Zou et al. 2014) which uses a (**Perceptron**)

Query Construction #1 - Issues

Construct a **SPARQL** query that reflects user question and gets the answer.
Semantic Gap: Issues with how the information is encoded in the KB. One cannot deduce how the information is stored from the question.

"Which countries are in the European Union?"
Could be encoded as:
dbr:Greece dbp:member dbr:European_Union
dbr:France dbp:member dbr:European_Union
or as:
dbr:Greece dct:subj dbc:Member_states_of_the_European_Union
dbr:France dct:subj dbc:Member_states_of_the_European_Union
How to search correctly

Query Construction #2

Approaches:

- Using templates
- Using information from the question analysis
- Using Semantic Parsers
- Using Machine Learning
- Using semantic information

Query Construction #3 - Templates

Templates with parts of the query to be filled, in general by triples.

- **QAKiS** (Cabrio et al. 2012) select queries with only one triple.
- **ISOFT** (Park, Shim, and Lee 2014) ASK over one triple, simple SELECT, COUNT and ORDER BY or FILTER.
- **PowerAqua** (Lopez et al. 2012) reduces the question to one or two triples ($<=$ 2 predicates).

Very restricted questions, language is too rich, disambiguity is key

Query Construction #4 - Question Analysis

Most systems get the form of the query in the question analysis and phrase mapping step.

- **Freya, Intui3** (Dima 2014) resources extracted in the phrase mapping step are combined into triples.
- **DEANNA** (Yahya et al. 2013) regex over POS tags in analysis step mapped to resources in phrase mapping step. ILP in dissambiguation step to get the triples.
- **gAnswer, QAnswer, RTV, SemGraphQA** (Zou et al. 2014, Ruseti et al. 2015, Giannone, Bellomaria, and Basili 2013, Beaumont, Grau, and Ligozat 2015) extract all the possible information from the dependency graph.

Query Construction #5 - Question Analysis

Xser (Xu, Feng, and Zhao 2014) 3 ML algorithms, two KB independant (on the question analysis phase), one KB dependant (on dissambiguation step)

- First algorithm: determines segments of the question corresponding to variables, properties, instances and classes.
- Second algorithm: find dependencies between phrases. (**Standford dependencies, PATTY**)
- Third algorithm: Dissambiguation with a **Structured Perceptron**

Query Construction #6 - Question Analysis

The problem with these methods is that they all **assume** that **is possible to deduce the structure of the SPARQL query from the structure of the question** without knowing how the knowledge is encoded in the KB.

Query Construction #7 - Semantic Parsing

Compose a grammar and use it to extract structure from the query.

- **G**rammatical **F**ramework (**GFMed** Marginean 2017)
- **F**eature-based **C**ontext **F**ree **G**rammar (**TR Discover**, Song et al. 2015)
- **C**ombinatorial **C**ategorial **G**rammar (Hakimov et al. 2015)
- **L**exical **T**ree **A**djoint **G**rammar (**TBSL** Unger et al. 2012, **BELA** Walter et al. 2012)

Query Construction #8 - Semantic Parsing

Question has to be well formulated. For each lexical item a corresponding semantic representation is needed. (ie married has to map with *dbo:spouse*). Learning corpus (Hakimov et al. 2015) or from POS tags (Unger et al. 2012). In general, **low recall**

Lexical item	Syntactic category	Semantic representation
Barack Obama	NP	$\mathtt{dbr:Barack\_Obama}$
is	$(S\backslash NP)/(S\backslash NP)$	$\lambda \mathtt{f}.\lambda \mathtt{x}.\mathtt{f}(\mathtt{x})$
married to	$(S\backslash NP)/NP$	$\lambda \mathtt{y}.\lambda \mathtt{x}.\mathtt{dbo:spouse}(\mathtt{x},\mathtt{y})$
Michelle Obama	NP	$\mathtt{dbr:Michelle\_Obama}$

Figure: Semantic parsed question *"Barack Obama is married to Michelle Obama"*

Query Construction #9 - Machine Learning

CASIA (He et al. 2014) (**low recall & precision**)

- Question Analysis step: extract features like position of a phrase and POS tags or the type of dependency in the dependency tree
- Phrase Mapping step: associate resources with phrase segments and extract more features
- Dissambiguation step: MLN with extracted features to find most probable relation between segments and most probable mapping. **retrained for each KB**

Query Construction #10 - Semantic information

SINA (Shekarpour et al. 2015), **POMELO** (Hamon et al. 2014), Zhang et al. 2016 do not rely on the syntactic features of the question, instead the whole process is done based on the KB, just with semantic information.

Advantages:

- **high recall, & precision**

Disadvantages:

- **computationally expensive**
- **does not respect user question syntax**. No difference between *"Who is the mother of Angela Merkel?"* and *"Angela Merkel is the mother of who?"*

Query Construction #11 - Summary

	Using templates	Using info. from the QA	Using Semantic Parsing	Using machine learning	Semantic information	Not generating SPARQL
BELA			x			
CASIA				x		
DEANNA		x				
FREyA		x				
gAnswer		x				
GFMed			x			
Hakimov et al.			x			
Intui2		x				
Intui3		x				
ISOFT	x					
POMELO					x	
PowerAqua	x					
QAKiS	x					
QAnswer		x				
RTV		x				
SemGraphQA		x				
SemSeK						x
SINA					x	
SWIP	x					
TBSL			x			
Treo						x
TR Discover			x			
UTQA	?					
Xser		x				
Zhang et al.					x	
Zhu et al.						x

Query Construction #12 - Summary

Which techniques to choose?

There is no clear way to do this. A good approach is to construct the query on the previous analysis steps, assuming structure can be extracted from the question.

Conclusions

There exists many techniques for each part:

- Smart balance between different techniques lead to best results
- Tendency goes to Machine Learning

When possible, maximize:

- KB independance (pluggability)
- Performance (real time, scalability)
- Extracted Knowledge (implicit, enriches the context/domain)

"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.

R. Beaumont, B. Grau, and A.-L. Ligozat. *SemGraphQA@QALD-5: LIMSI participation at QALD-5@CLEF. In: Working notes for CLEF 2015 conference*. en. CLEF, 2015.

References II

E. Cabrio et al. “QAKiS: an open domain QA system based on relational patterns”. en. In: *Proceedings of the 2012th international conference on posters demonstrations track-volume 914, CEUR-WS. org*. 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.

C. Dima. “Answering natural language questions with Intui3”. en. In: *Conference and labs of the evaluation forum (CLEF)*. 2014.

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.

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. Hakimov et al. "Applying semantic parsing to question answering over linked data: addressing the lexical gap". en. In: *Natural language processing and information systems*. Springer, 2015.

T. Hamon et al. *Description of the POMELO System for the Task 2 of QALD-2014*. en. In: CLEF (Working Notes, 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: *Semant Web* 3.249–265 (2012).

A. Marginean. "Question answering over biomedical linked data with grammatical framework". en. In: *Semant Web* 4.565–580 (2017).

References V

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. Park, H. Shim, and G.G. Lee. *ISOFT at QALD-4: semantic similarity-based question answering system over linked data*. en. In: CLEF, 2014.

S. Ruseti et al. *QAnswer-enhanced entity matching for question answering over linked data*. en. In: CLEF (Working Notes), CLEF, 2015.

References VI

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.

D. Song et al. *TR discover: a natural language interface for querying and analyzing interlinked datasets. In: The semantic web-ISWC*. en. Springer, 2015.

C. Unger et al. "Template-based question answering over RDF data". en. In: *Proceedings of the 21st international conference on world wide web, ACM*. 2012, pp. 639–648.

References VII

S. Walter et al. *Evaluation of a layered approach to question answering over linked data. In: The semantic web–ISWC*. da. Springer, 2012.

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.

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.

References VIII

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.

Y. Zhang et al. *Question answering over knowledge base with neural attention combining global knowledge information*. en. arXiv preprint arXiv:1606.00979. 2016.

C. Zhu et al. *Tian Y, Yu Y (2015) A graph traversal based approach to answer non-aggregation questions over DBpedia*. it. arXiv preprint arXiv:1510.04780.

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.