

Question Answering

General Principles and Variations

Tutorial: Constructing Question Answering Systems over
Knowledge Graphs

Dennis Diefenbach, Andreas Both, Pierre Maret

Definitions



Question Answering Systems

Systems that deliver the direct answer to a user's question based on a corpus of data or documents

Question Answering Systems

Systems that deliver the answer to a user's question based on a corpus of data or documents

Information Retrieval, Search Engines

IR: Searching for documents from a corpus.

Search engine: IR engine where the corpus is the Web. The results are web pages.

Characteristics of QA Systems

The general principles

- General process
 - NL question analysis
 - Translation into formal queries
 - Execution
 - Result processing and display
 - Combination of techniques
 - Pattern matching
 - Natural Language Processing
 - Semantic Web
 - Machine learning
 - ...
-

Several Variations

1. Corpus: document, data, hybrid
 2. Corpus/Domain specific or free
 3. One-shot question or conversation
 4. Single text field or complex UI
 5. Core processing approach
-

Variation 1: Type of corpus

TEXT

QA over Text

Also called Information Retrieval-based QA systems

Specificity: Find the parts of text that contain the answer, and extract it.

DATA

QA over Data

Structured data: Databases, Linked data

Specificity: Find the field values that answer to the question

HYBRID

QA over an hybrid corpus

Variation 2: Corpus/Domain specific or free

Corpus specific

Advantages: Vocabulary, data structure, training data are known

Drawbacks: No portability

Domain specific

Advantages: Resources are available (vocabulary, thesaurus)

Drawbacks: No portability (cost of the collecting resources)

Corpus/Domain free

Advantages: No predefined lexical resources required, potential extendability and portability to other fields

Drawbacks: Results harder to achieve

Variation 3: One-shot questions or conversation

One-shot questions

The query is built with the NL question only

Each question is a new one.

Conversational system

Chatbot systems

The query is enriched with a context

Context: 'n' previous queries and answers

System driven conversation: The system leads the conversation to resolve ambiguities (or achieve goals)

User driven conversation: The system only answers to successive user's questions.

Variation 4: Single text field or richer UI

Rich UI facilitates the query process

Typed text fields

Check boxes organized in facets

Drawbacks

Domain specific

More complex for the users (i.e. fields to fill, selection in a list of possible values)

Variation 5: Core processing approach

All approaches combine with different emphasis several techniques such as NLP, ML, SW.

Use of Lexico-Syntactic Patterns

Supervised-learning approach: A model is learned to associate questions to patterns

Drawbacks: bad performances for keyword and malformed questions; language dependant; requires training data

Use of Semantic Web Resources

Graph algorithm approach: Retrieval of resources and disambiguation

Drawbacks: NL semantics may be lost, size of the candidate resource list

Criteria for the evaluation of QA Systems

- Precision and recall
- Response time and scalability
- Portability
 - Other datasets
 - Other languages

Benchmarks



Datasets	WebQuestions	SimpleQuestions	QALD 1 to 9
Nb of questions	5.810	108.442	50 to 250
Year of publication	2013	2015	2011 to 2018
Types of relations implied	Reified statements (97%)	Single statements (1 triple)	Up to 3 binary relations
Comments			Often need modifiers (ORDER BY, COUNT)
Language	English	English	Multilingual (since 5)

Industrial application

—

Industrial applications

	Wolfram Alfa	SIRI	Google assistant	QAnswer
Company	Wolfram Research	Apple	Google	QAnswer
Database for public access	Proprietary	Proprietary	Proprietary	Wikidata, Dbpedia, Openstreet map...
Start	2009	2011	2016	2018
Languages	English	20+	20+	European languages and Arabic, Chinese, Japanese

Industrial applications

	Wolfram Alfa	SIRI	Google assistant	QAnswer
Type of corpus	Data	Data	Data	Data
Corpus/Domain specific or free	Domain free (fields of specialities)	Domain free	Domain free	Domain free
One-shot question or conversation	One-shot question	Conversation	Conversation	One-shot question
Single text field or complex UI	Single text field (+field selection)	Single text field (+ voice)	Single text field (+ voice)	Single text field (Voice)
Core processing approach	Unknown	Unknown	Unknown	Semantic Web

The end

Thanks for your
attention