Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 73 (2015) 366 – 375

Procedia Computer Science

[S1]The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

Question Answering Systems: Survey and Trends

Abdelghani BOUZIANE[a], Djelloul BOUCHIHA[a], Noureddine DOUMI[b] and Mimoun MALKI[c]

[a]Ctr Univ Naama, Inst. Sciences and Technologies, Dept. Mathematics and Computer Science, Algeria
{Ghani.ab1; bouchiha.dj}@gmail.com

[b]University Dr. Tahar Moulay of Saida
noureddine.doumi@univ-saida.dz

[c]EEDIS Laboratory, Djillali Liabes University of Sidi Bel Abbes, Algeria
Malki.Mimoun@univ-sba.dz

Abstract

The need to query information content available in various formats including structured and unstructured data (text in natural language, semi-structured Web documents, structured RDF data in the semantic Web, etc.) has become increasingly important. Thus, Question Answering Systems (QAS) are essential to satisfy this need. QAS aim at satisfying users who are looking to answer a specific question in natural language. In this paper we survey various QAS. We give also statistics and analysis. This can clear the way and help researchers to choose the appropriate solution to their issue. They can see the insufficiency, so that they can propose new systems for complex queries. They can also adapt or reuse QAS techniques for specific research issues.

Peer-review under responsibility of organizing committee of the International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

Keywords: Question Answering System (QAS), Natural Language Processing (NLP), Information Retrieval, SPARQL, Semantic Web;

Peer-review under responsibility of organizing committee of the International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)
doi:10.1016/j.procs.2015.12.005

1. Introduction

The rapid increase in massive information storage and the popularity of using the Web allow researchers to store data and make them available to the public. However, the exploration of this large amount of data makes finding information a complex and expensive task in terms of time. This difficulty has motivated the development of new adapted research tools, such as Question Answering Systems.

In fact, this kind of system allows the user to ask a question in natural language (NL) and return the right answer to his question instead of a set of documents deemed relevant, as is the case for engines research.

However, for Question Answering Systems dedicated to manipulate text and Web documents, the structure of the required information affects the accuracy of these systems. QAS are most effective to interact with structured knowledge bases.

Due to the importance QAS, Other surveys are available in the literature, like [1] and [2]. In our survey paper:

- We count Question Answering Systems and analyze the propositions according to different points of view,
- We refresh existing surveys by adding recent works,
- Motivated by our ongoing project, titled QAS for Arabic Linked Data, we give a classification based, in particularly, on language and data-structure dimensions.
- Statistics presented through graphical histograms give clear view to researchers working in this field.

The rest of the paper is organized as follows: Section 2 describes some notions related to the discussed issue in the paper. Section 3 cites and classifies Question Answering Systems, Section 4 provides statistics on the QAS. In Section 5, we describe the project that we are working on. Finally, Section 5 concludes our work.

2. Background

Many notions have to be learned before counting works on Question Answering Systems.

2.1. What is Question Answering System?

Many definitions are available in the literature:

“For human-computer interaction, natural language is the best information access mechanism for humans. Hence, Question Answering Systems (QAS) have special significance and advantages over search engines and are considered to be the ultimate goal of semantic Web research for user’s information needs” [3].

“Question Answering on the Web is moving beyond the stage where users simply type a query and retrieve a

ranked ordering of appropriate Web pages. Users and analysts want targeted answers to their questions without extraneous information“ [4].

In this paper, we focus in particularly on QAS dedicated to the Web of documents and the Web of data.

2.2. QAS for Web of documents and text

In Information Retrieval (IR) and Natural Language Processing (NLP), Question Answering (QA) is the task of automatically providing an answer for a question asked by a human in natural language. QA as a task can be divided into three main distinct subtasks, which are: Question Analysis, Document Retrieval and Answer Extraction [5] (see Fig. 1). Most Question Answering Systems follow these three subtasks. However, they may differ in how they implement every subtask.

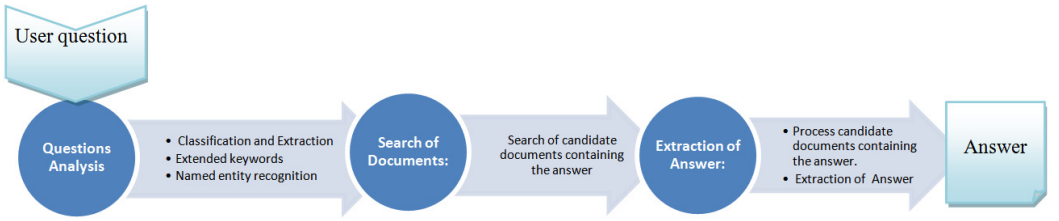

Fig 1: Subtasks of QAS dedicated to Web of documents

The Question Answering issue deals with the natural language processing for interfacing the QAS at the side of users who ask many types of questions. In particularly, Factoid questions are those asked mainly about Named Entity (NE), using for example the words: When, Where, How much/many, Who, and What, which ask respectively about date/time, place, person, and organization. The Second type is the questions that ask about the definition of term or concept. Questions that use the words "Why" or "How" are another type that is hard to answer and there are very little if any attempts done to answer this type of questions.

2.3. QAS for the Web of data

The goal of QA Systems, as defined by [6], is to allow users to ask questions in Natural Language (NL), using their own terminology, and receive a concise answer. For QAS dedicated to the Web of data, User asks question in natural language. The process starts by linguistically analyzing (dependency graphs using a syntactical parser with a step of named entities recognition NER). The next step is to classify the question according to one defined question category. The SPARQL query is generated in two steps (linguistically analyze and question classification). An external ontology resource can be used for matching items generated in the process. Finally, when the SPARQL query is generated, the interrogation of the Linked Data is done, and generates the exact answer of the user question.

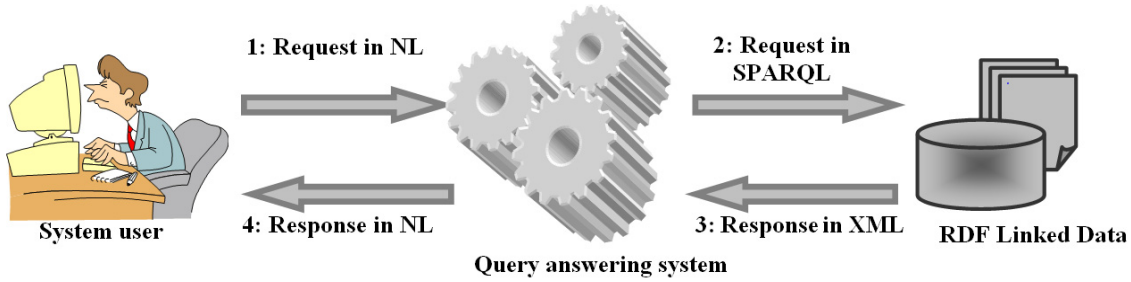

Fig 2: The scenario of interacting with a Question Answering System dedicated to Linked Data.

3. Question Answering Systems

Question Answering Systems present a good solution for querying unstructured and structured information. This is why a large number of QA systems have been developed to various languages. Some languages, such as Latin, in particularly English, are better served than others, such as Arabic and Semitic language in general. This might be related to the language features and the maturity of research in the countries speaking it [7].

Next we present a survey of QAS, with different sources: structured databases, unstructured free text and precompiled semantic knowledge bases.

3.1. QAS for Latin languages

Due to the popularity, importance and features of the English language, tens of QA Systems are available since 1960, like BASEBAL system [8]. The current trend is moving towards Lined Data. Next, we highlight some of the most prominent work on this area.

Table 1: QAS features and techniques for Latin languages

QAS features		**appear in**	**Used techniques**
NLIDB:(Natural Language Interfaces to Databases)		PRECISE [9]	Identifying classes of questions
		The formal semantic approach [10]	Intermediate representation language
		MASQUE/SQL [11]	Portable NL front end to SQL databases
		BASEBALL [8] [12]	Specific domain Systems
Open Domain Question Answering	Document-based Question	[13], [14], [16], LASSO [15]	Deep linguistic analysis and iterative strategy
		FALCON [17]	Hierarchies of question types based on the types of answers sought

over text	Answering	DIMAP [18]	Semantic categories of answers are mapped into categories covered by a NE Recognizer. When the answer type is identified, it is mapped into answer taxonomy, where the top categories are connected to several word classes from WordNet.
	Question Answering On the Web	Mulder [19]	Extracting "semantic relation triples" after the document is parsed, converting the document into triples.
		FAQ Finder [20]	QA System for factual questions over the Web
		QALC [21]	Statistical or semantic similarities
		QRISTAL [22]	Provides answers to English factoid questions based on syntactic and semantic analysis
		WebQA [23]	Based on named entities' recognition, and conceptual and thematic analysis
		Ask.com [24]	Using the template-mapping technique to define the question type clustering technique to extract multiple answer blocks

3.2. Ontology based Question Answering Systems

Ontology based QA Systems take queries expressed in NL and a given ontology as input, and return answers drawn from one or more KBs that subscribe to the ontology. Therefore, they do not require the user to learn the vocabulary or the structure of the ontology.

Ontology based QA Systems vary on two main aspects: (i) the degree of domain customization they require, which correlates with their retrieval performance, and (ii) the subset of NL they are able to understand (full grammar-based NL, controlled or guided NL, pattern based).

Table 2: Ontology based QAS features and techniques

QAS features	**appear in**	**Used techniques**
QA Systems based on ontologies	AquaLog [25]	Allows the user to choose an ontology and then ask NL queries with respect to the universe of discourse covered by the ontology
	PowerAqua [26]	QAS focusing on querying multiple semantic Web resources
	QACID [27]	Relies on an ontology, a collection of user queries, and an entailment engine that associates new queries to a cluster of existing queries.
	ORAKEL [28]	Translates factual wh-queries into F-logic or SPARQL and evaluates them with respect to a given KB
	GINSENG [29]	Controls user's input via a fixed vocabulary and predefined sentence structures through menu-based options
	PANTO [30]	Portable NLI that takes a NL question as input and executes a corresponding SPARQL query on a given ontology

		model
	FREyA [31]	Providing improvements with respect to a deeper understanding of a question's semantic meaning
	QAKIS [32]	Technique for matching NL fragments and textual patterns, auto-collected from Wikipedia
	SPARQL2NL [33]	In the side of converting a SPAQL query into natural language.
	SWIP [34]	The processing of the NL query is based on the use of the pivot query: from the NL user query into a pivot query, and the formalization of this pivot query.
	Pythia [35]	Using ontology in the process of interpretation of user query
	SQUALL [36]	Using a controlled natural language for translation to SPAQL query
	TBSL [37] LODQA [38]	The user question is transformed to a template query (a mirror template). From the NL query to generate the SPAQL query using the template model.
	DeepQA IBM Watson's system [39]	Using unstructured and structured data (RDF format) to extract and score evidence
	CASIA [40]	A Markov Logic Networks algorithm is used for learning a joint model, for detecting phrases and for mapping semantic Items. For these phases, the semantic items are grouping into a graph.

4. Question Answering performance

In this section we present statistics about two types of QA systems: Ontology based QAS and Text based QAS:

4.1. Ontology based QA Systems

To show the performance of the ontology based QA Systems we looked at the evaluation results carried out in the literature, notably those summarized in the survey paper [5]. Then we establish the histogram of the following figure.

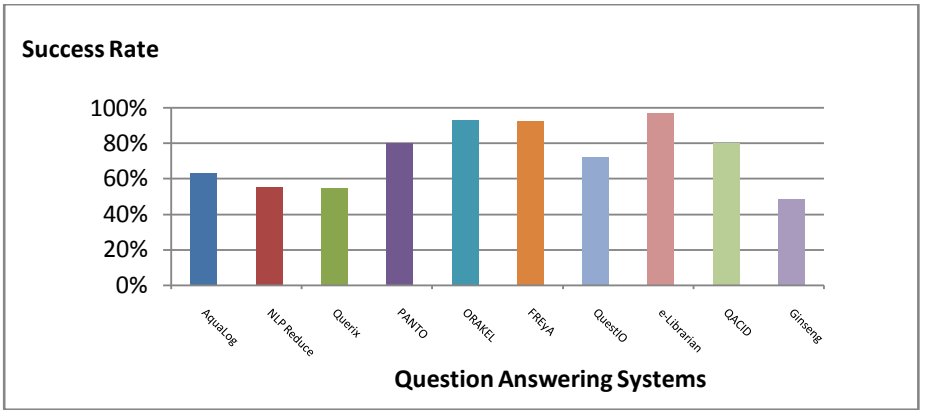

Fig 3: Performance results of the ontology-based Question Answering Systems.

Performance of the Ontology based QAS is represented by the success rate (correct answers to questions) in the graph above.

We found that the success rate of these QA systems varies between 49% and 89%. These results depend on two criteria: (1) the algorithms and methods of natural language processing used by the system, and (2) the specified domain to be questioned by the system.

4.2. Text based QA Systems

To evaluate the Texts based QA Systems we looked at the results given in the Question Answering for Machine Reading (QA4MRE), the Main Task at the 2013 Cross Language Evaluation Forum [41].

The QA4MRE focuses on the reading of single documents and the identification of the correct and NoA answers to a set of questions, over the two years 2012 and 2013. NoA means that the system decided not to answer the question.

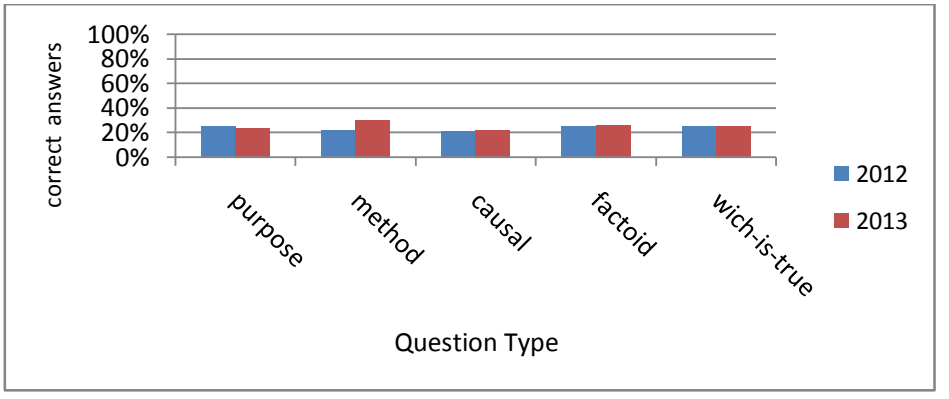

Fig 4: Shows the percentage of correct answers for different question types in the 2012/2013 versions of the QA4MRE challenge.

Finally, we can say that QAS reliability increases in direct proportion to percentage of correct answers, and is in inverse proportion to percentage of NoA answers.

5. What are we planning for?

Our project is to implement Question Answering System to explore Linked Data. The system user can formulate his request with Arabic natural language. The system converts then the request into SPARQL request to interrogate Arabic RDF Linked Data, and finally returns the results to the user.

The global process of the system is illustrate in fig 8 many modules such as the Arabic Natural Language module, semiformal query module and ANL generation module will be implemented to achieve our goal.

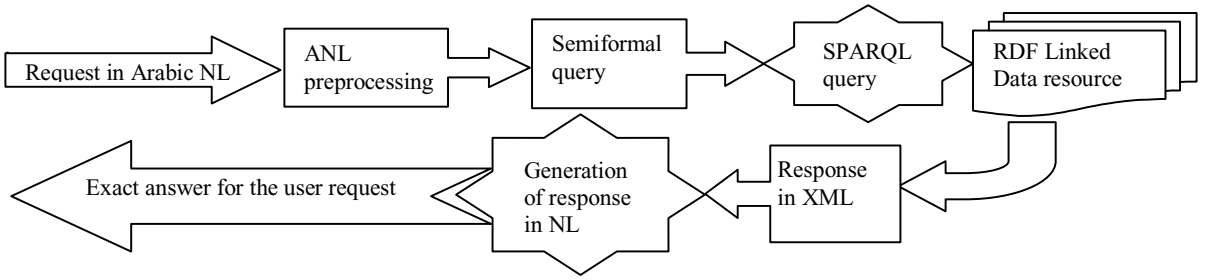

Fig 5: Steps and modules of The Arabic Query Answering System

Many problems have to be solved. Natural Language Processing (NLP) techniques can be used to convert the user request from NL into SPARQL. Then, other APIs can be used to return the results to the user.

6. Conclusion

A Question Answering System aims at giving precise answers to users' questions introduced in natural language. The purpose of this paper is to cite and classify many QAS. This can clear the way for researchers in this domain. They can choose the appropriate system to their problem. They can also see the shortcomings and correct them, or propose new QA Systems.

It is important to note that one of the most important features of QA Systems is their ability to provide exact answers, because different sources are the target of these systems. Then, the user asks a question using a natural language without knowing the structure of the sources to be queried. Some languages are better served than others, due to the maturity of research in the countries speaking this language. So the research in natural language processing is primordial for developing Question Answering Systems for unstructured and structured data. The Arabic semantic Web is very far from the development of Arabic Question Answering System over semantic Web, which is our ultimate goal to achieve.

References

1. Allam A. and Haggag M., "The Question Answering Systems: A Survey". International Journal of Research and Reviews in Information Sciences (IJRRIS), Vol. 2, No. 3, September (2012)
2. Kalaivani S. and Duraiswamy K., "Comparison of Question Answering Systems Based on Ontology and Semantic Web in Different Environment". Journal of Computer Science, Vol. 8, No. 9 (2012)
3. Trotman A., Geva S. and Kamps J., "Report on the SIGIR 2007 workshop on focused retrieval". SIGIR Forum, Vol. 41, No. 2, pp.97–103. (2007)
4. Hagen P.R., Manning H. and Paul Y., "Must Search Stink?". Forrester Research, TechStrategy Briefing, June (2000).
5. Lopez V., Uren V., Sabou M. and Motta E., "Is Question Answering fit for the Semantic Web?: a Survey". Universität Bielefeld, Germany, (2011)
6. Hirschman L. and Gaizauskas R., "Natural Language Question Answering: The View from here". Natural Language Engineering, Special Issue on Question Answering". 7(4): 275-300. Cambridge University Press. (2001)
7. Kurdi H., Alkhaider S. and Alfaifi N., "DEVELOPEMNT AND EVALUATION OF A WEB BASED QUESTION ANSWERING SYSTEM FOR ARABIC LANGUAG". International Journal on Natural Language Computing (IJNLC) Vol. 3, No.2, April (2014)
8. Green B.F., Wolf A. K., Chomsky C. and Laughery K., "BASEBALL: An automatic question answerer". Proceedings Western Joint Computer Conference, 19:207-216. McGraw-Hill (1961)
9. Popescu A. M., Etzioni O. and Kautz H. A., "Towards a Theory of Natural Language Interfaces to Databases". In Proc. of the International Conference on Intelligent User Interfaces, p. 149-157. ACM Press. (2003)
10. De Roeck A N., Fox C J., Lowden B G T, Turner R., Walls B., "A Natural Language System Based on Formal Semantics". In Proc. of the International Conference on Current Issues in Computational Linguistics, Penang, Malaysia. (1991)
11. Androutsopoulos I., Ritchie G.D. and Thanisch P., "Natural Language Interfaces to Databases - An Introduction. Natural Language Engineering". 1(1): 29-81.Cambridge University Press. (1995)
12. Woods W., "Progress in natural language under-standing - an application to lunar geology". In Proc. of the American Federation of Information Processing Societies (AFIPS), 42: 441-450. AFIPS Press. (1973)
13. Hovy E. H., Gerber L., Hermjakob U., Junk M. and Lin C.Y., "Question Answering in Webclopedia.". In Proc.of the Ninth Text Retrieval Conference (TREC-9). NIST, p.655-664. (2000)
14. Wu M., Zheng X., Duan M., Liu T. and Strzalkowski T., "Question Answering by Pattern Matching, WebProofing, Semantic Form Proofing". NIST Special Publication: The Twelfth Text REtrieval Conference (TREC),p. 500-255. (2003)
15. Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Girju R. and Rus V., "LASSO: A Tool for Surfing the Answer Net". In Proc. of the Text Retrieval Conference (TREC-8). (1999)
16. Srihari K., Li W. and Li X., "Information Extraction Supported Question- Answering. In Advances in Open Domain Question Answering". Kluwer Academic Publishers. (2004)
17. Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R, Girju R., Rus V. and Morarescu P., "Falcon - Boosting Knowledge for Answer Engines". In Proc. of the 9th Text Retrieval Conference (Trec-9), p.479-488. (2000)
18. Litkowski, K. C., "Syntactic Clues and Lexical Resources in Question-Answering". The Ninth Text REtrieval Conferenence (TREC-9), NIST Special Publication 500-249. (2001)
19. Kwok C., Etzioni O. and Weld D., "Scaling question answering to the Web". In Proc. of the 10th International Conference on World Wide Web, p.150-161, Hong Kong, China. ACM (2001)
20. Burke R. D., Hammond K. J.and Kulyukin V., "Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder system". In Proc. of the World Wide Web Internet and Web Information Systems, 18(TR-97-05): 57-66. Department of Computer Science, University of Chicago. (1997)
21. Ferret O., Grau B., Hurault-Plantet M., "Finding an answer based on the recognition of the issue focus". In Proceedings of TREC-10. (2000).
22. Laurent D., Séguéla P. and NègreCross S., "Lingual Question Answering using QRISTAL for CLEF 2006". Lecture Notes in Computer Science, Vol. 4730, 2007, pp. 339-350.
23. Parthasarathy S. and Chen J., "A Web-based Question Answering System for Effective e-Learning". In

Proceedings ofIEEE International Conference on Advanced Learning Technologies, 2007, pp. 142-146. (2007)
24. Ask.com, checked Aug. 2nd, (2013).
25. Lopez V., Uren V., Motta E. and Pasin M., ”AquaLog: An ontology-driven question answering system for organizational semantic intranets”. Journal of Web Semantics: Science Service and Agents on the World Wide Web, 5(2): 72-105. (2007)
26. Lopez V., Fernández M., Motta E., and Stieler N., “PowerAqua: supporting users in querying and exploring the Semantic Web”. Semantic Web 3(3), 249–265 (2012)
27. Fernandez O., Izquierdo R., Ferrandez S., Vicedo J. L., “Addressing Ontology-based question answering with collections of user queries“. Information Processing and Management, 45 (2): 175-188. Elsevier.(2009)
28. Cimiano P., Haase P. and Heizmann J., “Porting Natural Language Interfaces between Domains An Experimental User Study with the ORAKEL System“. In Chin, D. N., Zhou, M. X., Lau, T. S. and Puerta A. R., editors. In Proc.of the International Conference on Intelligent User Interfaces, p. 180-189, Gran Canaria, Spain. ACM. (2007)
29. Bernstein A., Kauffmann E., Kaiser C. and Kiefer C., ”Ginseng: A Guided Input Natural Language Search Engine”. In Proc. of the 15th workshop on Information Technologies and Systems (WITS 2005), p. 45-50. MV-Wissenschaft,Münster. (2006)
30. Wang C., Xiong M., Zhou Q. and Yu Y., “PANTO: A portable Natural Language Interface to Ontologies”. In Franconi, E., Kifer, M., May, W., editors. In Proc. of the 4th European Semantic Web Conference, p.473-487, Innsbruck,Austria. Springer Verlag. (2007)
31. Damljanovic D., Agatonovic M. and Cunningham H,, ”Natural Language interface to ontologies: Combining syntactic analysis and ontology-based lookup through the user interaction”. In Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L. and Tudorache, T., editors. In Proc. of the European Semantic Web Conference, Heraklion, Greece. Springer Verlag. (2010)
32. Cabrio E., Cojan J., Aprosio A. P., Magnini B., Lavelli A. and Gandon F., “QAKiS: an open domain QA system based on relational patterns”. In Proceedings of the ISWC 2012 Posters & Demonstrations Track. CEUR Workshop Proceedings, vol. 914 (2012)
33. Axel-Cyrille N. N., Bühmann L. and Unger C., “Sorry, I don’t speak SPARQL – Translating SPARQL Queries into Natural Language”. International World Wide Web Conference Committee (IW3C2).WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil. ACM 978-1-4503-2035-1/13/05. (2013)
34. Pradel C., Haemmerlé O. and Hernandez N., “Swip: A Natural Language to SPARQL Interface Implemented with SPARQL”, ICCS 2014, LNAI 8577, pp. 260–274, 2014.Springer International Publishing Switzerland (2014)
35. Unger C. and Cimiano P., “Pythia: Compositional meaning construction for ontology based question answering on the semantic web”. In: Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 153–160. Springer, Heidelberg (2011)
36. Ferré S., “SQUALL: A Controlled Natural Language as Expressive as SPARQL 1.1”. NLDB 2013, LNCS 7934, pp. 114–125, 2013. Springer-Verlag Berlin Heidelberg (2013)
37. Unger C., Bühmann L., Lehmann J., Ngomo A. C. N., Gerber D. and Cimiano P., ”Template-based question answering over RDF data”. In Proceedings of the 21st International Conference on World Wide Web, pp. 639–648. ACM (2012)
38. Kim J. D. and Cohen K. B., ”Natural language query processing for SPARQL generation: A prototype system for SNOMEDCT”. In Proceedings of BioLINK SIG (2013)
39. Kalyanpur A., et al., “Structured data and inference in DeepQA”. IBM Journal of Research & Development 56(3/4) (2012)
40. Shizhu H., Yuanzhe Z., Liu K. and Zhao J., “CASIA@V2: A MLN-based question answering system over linked data”. In CLEF 2014 Working Notes Papers, (2014).
41. Berners-Lee T., Hendler J. and Lasilla O., “The Semantic Web”. Scientific American. (2001)