Ontology-Based Question Analysis Method

Ghada Besbes[1], Hajer Baazaoui-Zghal[1], and Antonio Moreno[2]

[1] Riadi-GDL, ENSI Campus Universitaire de la Manouba, Tunis, Tunisie
ghada.besbes@gmail.com, hajer.baazaouizghal@riadi.rnu.tn
[2] ITAKA Research Group, Departament d'Enginyeria Informatica i Matematiques, Universitat Rovira i Virgili, Av. Paisos Catalans, 26. 43007, Tarragona, Spain
antonio.moreno@urv.cat

Abstract. Question analysis is a central component of Question Answering systems. In this paper we propose a new method for question analysis based on ontologies (*QAnalOnto*). *QAnalOnto* relies on four main components: (1) Lexical and syntactic analysis, (2) Question graph construction, (3) Query reformulation and (4) Search for similar questions. Our contribution consists on the representation of generic structures of questions and results by using typed attributed graphs and on the integration of domain ontologies and lexico-syntactic patterns for query reformulation. Some preliminary tests have shown that the proposed method improves the quality of the retrieved documents and the search of previous similar questions.

Keywords: Question-Answering systems, ontology, lexico-syntactic patterns, typed attributed graphs.

1 Introduction

With the rapid growth of the amount of online electronic documents, the classic search techniques based on keywords have become inadequate. Question Answering systems are considered as advanced information retrieval systems, allowing the user to ask a question in natural language (NL) and returning the precise answer instead of a set of documents. The search process in a Question Answering system is composed of three main steps: question analysis, document search and answer extraction from relevant documents. Generally, Question Answering (QA) systems aim at providing answers to NL questions in an open domain context and can provide a solution to the problem of response accuracy. This requirement has motivated researchers in the QA field to incorporate knowledge-processing components such as semantic representation, ontologies, reasoning and inference engines. Our work hypothesis is that, if the user starts with a well-formulated question, answers will be more relevant; this is why, in this work, we focus on question analysis. So, the aim of this paper is to design and implement a new method dedicated to question analysis in a QA system. Indeed, our goals consist on improving the representation of the question's structure by using typed attributed graphs and improving the results of query reformulation by using domain ontologies and lexico-syntactic patterns.

H.L. Larsen et al. (Eds.): FQAS 2013, LNAI 8132, pp. 100–111, 2013.

In this method, first of all, lexical and syntactic analyses are applied to the user's question. Second, a question graph, containing all the information about the question, is constructed based on a generic question graph using knowledge from WordNet and from a question ontology. Then the question is reformulated based on lexico-syntactic patterns and the domain knowledge represented in an ontology. Finally, the method stores the question graph and the reformulated question in a question base in order to extract analysis results for similar questions later. Our method is dedicated to QA systems as it deals with NL queries asked in a question form, considered as a particular case of information retrieval systems. The evaluation is conducted using information retrieval metrics such as precision and MAP.

The remaining of this paper is organized as follows. Section 2 presents an overview of works related to question analysis techniques. Section 3 describes our method of question analysis based on ontologies. Section 4 presents and discusses some experimental results of our proposal. Finally, section 5 concludes and proposes directions for future research.

2 Related Works

The question analysis component is the first step of the search process in a question-answering system. This analysis aims to determine the question's structure as well as the significant information (expected answer type, terms' grammatical functions, etc.) that are considered as clues for identifying the precise answer. Question analysis methods can be classified depending on their level of linguistic analysis: (i) **Lexical analysis:** The lexical level of NL processing is centered on the concept of a word, and the techniques used for lexical analysis are generally a pre-treatment for the following analysis. The most used techniques are the following: tokenization (division of the question into words) and keyword extraction [1], lemmatization [2](considering the root to group words of the same family) and removing stop words [1] (the elimination of common words that do not affect the meaning of the question to reduce the number of words to be analyzed). (ii) **Syntactic analysis:** Information extracted from the question analysis component is the basis for answer extraction. This component constructs a representation of the question, which differs from one system to the other and contains various types of information and knowledge. The purpose of this analysis is to preserve the syntactic structure of the question by exploiting the syntactic functions of words in the questions [3]. Question-answering systems use different techniques of NL processing, including the following: Part-of-speech tagging or POS tagging [4] (giving each word a tag that represents information about its class and morphological features), named entity recognition (identifying objects as classes that can be categorized as locations, quantity, names, etc.) and the use of a syntactic parser. (iii) **Semantic analysis:** In some question-answering systems, analyzing the question goes beyond vocabulary and syntax up to semantics and query reformulation. This phase includes the extraction of semantic relations between the question words [5] to make a semantic representation as in the Javelin system [6]. The purpose of semantic analysis is

to detect and represent semantic knowledge in order to use it for inference or matching when extracting the answer. To do this, several systems rely on semantic techniques in order to have a better analysis of the question. In the case of query reformulation and enrichment, most systems use tools and semantic knowledge such as WordNet [7] or ontologies to extract other semantic forms for the question keywords. In fact, ontology-based question-answering systems such as QuestIO [8], AquaLog [9] and QASYO [12] use an internal representation of knowledge in the form of an ontology. The purpose of using an ontology as a knowledge representation is either to extract the answer directly as in Querix [10] or to reformulate the query by rewriting the user's question using the ontology concepts.

In general, the purpose of question analysis is to collect information on the subject of the question, to represent it and to formally submit a request to the search engine. The previous study allows identifying the following limits on the different levels of linguistic analysis: (i) **Lexical analysis:** Question analysis in many question-answering systems is reduced to the lexical analysis, and extracted keywords are used as search queries for the information retrieval system without any reformulation. This method does not represent the question and does not extract the terms' grammatical functions. (ii) **Syntactic analysis:** with only a syntactic analysis, the query reformulation problem is still not resolved. In addition, the question's representation has only the terms used in it and their morpho-syntactic classes; therefore, it does not represent the question's semantic knowledge.(iii) **Semantic analysis:** Query reformulation at this level focuses only on retrieving potentially relevant documents, not answer-bearing ones.

Our main objective is to improve the question analysis component in QA systems in order to improve their performance. During the study of the state of the art we identified the following items to address: finding similar questions from a question base, representing analyzed questions and reformulating the queries.

1. The process of finding similar questions in a question base is a computationally expensive, and most similarity measures are designed to deal with concepts not with questions. We therefore applied a filtering on the question base in order to lighten the process and we combined statistic and semantic similarity measures suitable for questions.
2. During the question analysis process, we are confronted with the problems of determining its structure and the lack of expressiveness of representation formalisms that do not respect the granularity of the concepts used in the question. Therefore, we used a generic graph (in fact, a typed attributed graph) to represent the structure of the question
3. Query reformulation is not rich enough. It lacks external knowledge such as ontologies to bring new concepts and terms. It is also oriented towards relevant documents not answer-bearing ones. Through our method, we tried to solve the problems of query reformulation by using a domain ontology combined with lexico-syntactic patterns.

3 Question Analysis Method

The proposed method relies on four main components: (1) Lexical and syntactic analysis, (2) Question graph construction,(3) Query reformulation and (4) Search for similar questions. These components will be detailed in the following sections.

3.1 Proposed Method's Description

Figure 1 provides a general view of the proposed method for analyzing NL questions. The goal is to identify all the terms of a question and their grammatical functions in the question and to obtain useful information for the answer's extraction. First, the user submits a question in NL. The method performs a lexical and syntactic analysis (1). The syntactic analysis is based on POS tagging in order to identify the grammatical morpho-syntactic class for each term used in the question. These results are interpreted by the question ontology, synonyms for each term are extracted from WordNet and the structure of the question is defined using the generic graph that contains all the general structures of questions. The method builds a typed attributed graph (2) that contains all the information available in the question. Then, the question is reformulated (3) using lexico-syntactic patterns and the concepts of a domain ontology. The patterns required in this reformulation process are retrieved from the question ontology. Concepts that are semantically related to the terms of the question are extracted from a domain ontology to enrich the question. Using a question base, all questions are recorded along with their analysis results, that is to say, the typed attributed graph of the question and the result of query reformulation. Thus the method can search for similar questions (4) and the user has the option to extract directly the results of analysis of a stored similar question. The output of the method is a reformulated query ready to be submitted to a search engine and a set of useful, well-structured questions that will be used by the answer extraction component.

3.2 Lexical and Syntactic Analysis Component

The first step is lexical analysis which includes the following two processes:

- Tokenization: It's the division of the text into words that can be managed in the next steps of the analysis.
- Lemmatization: This process considers the root of a word. For example, all verbs are reduced to the infinitive (eaten, ate -> eat), plural nouns are reduced to singular, etc. In this way, a search using any of the word's variants will lead to the same result.

The second step is syntactic analysis. At this level of analysis, POS tagging is applied to the question. This is the process of associating a tag to each word in the question that represents information about its class and morphological features.

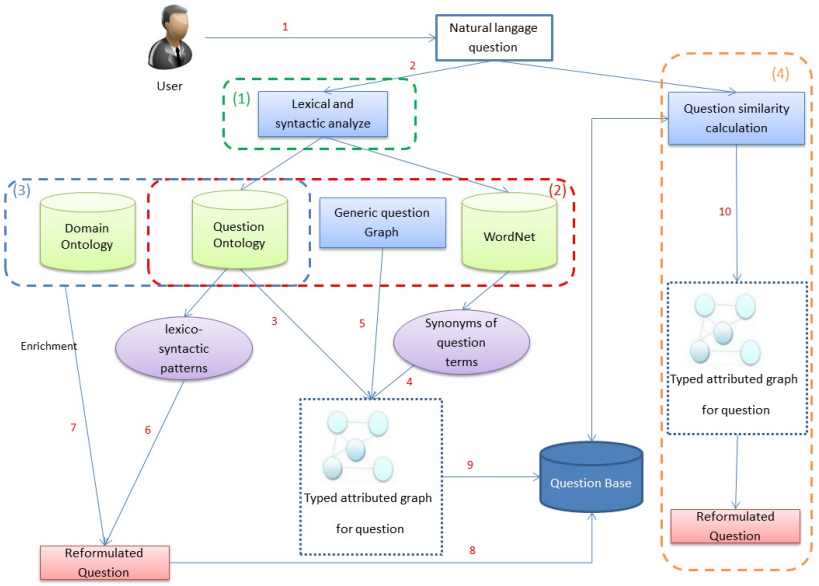

Fig. 1. General architecture of QAnalOnto

3.3 Question Graph Construction Component

The question graph is a representation of the user's question in an intuitive and understandable form that contains all the information included in the question necessary to search for its answer.

Generic Question Graph. The main advantage of using graphs resides in its capability to represent relations, even multiple ones, between objects. The generic question graph contains all forms of predefined questions. It is used to identify the question's structure. It is a typed attributed graph. This type of graph is a pair $(NG; EG)$where NG is a set of attributed nodes and EG is a set of attributed edges. An attributed node $n \in NG = (Tn, AVn)$ has a type Tn and a set of attribute values AVn. An attributed edge $e \in EG = (Te; AVe; Oe, De)$ has a type Te, a set of attribute values AVe, an attributed node that represents the origin of the edge Oe and an attributed node that represents the destination of the edge De.

Figure 2 shows a generic graph of a simple question: WH + Verb + Subject.

The nodes "WH","Verb" and "Subject" are subgraphs composed of "Term" nodes that represent, respectively, the kind of a WH-question, the main verb in the question and its subject.

A node "Term" (Ti) is the smallest conceptual unit representing a term in a question. The node "Term" consists of the following attributes: type ("Term"), value (question term), POS tag, lemma, category (WH, verb, subject) and synonyms (extracted from WordNet).

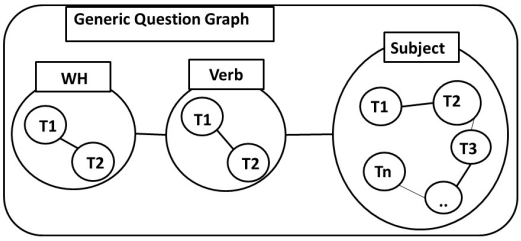

Fig. 2. Example of generic question graph

Each node "WH", "Verb" and "Subject" is itself a typed attributed graph $C = (TC; RTC)$ where TC is the list of attributed "Term" nodes and RTC is a set of typed edges between terms which represent the relation "followed_by" which specifies the order of the different terms of the question.

Construction Steps. In the first component, we performed a lexical and syntactic analysis on the question in order to extract the terms used in the question and their tags. Using these results, the system constructs the question graph. The construction process is divided on three steps:

1. Detection of the question's structure: Using the parsed question and the question ontology we can extract the question's structure from the generic question graph (that contains the structures of all types of questions allowed in the system). The system passes the parsed question by the question ontology in order to interpret the tags and determine the answer type. The question ontology is a manually constructed ontology that contains all the tags classified by category, so, tags are recognized and returned to the generic graph to identify and extract the question's structure. In fact, in the question ontology each kind of question has different answer types. From the results of the tagging, the ontology defines the expected answer type for the question. The ontology also contains lexico syntactic patterns for each type of question, that can be used to reformulate it.
   Part of the question ontology focused on the question "where" is represented in Figure 3.

   The ellipsis boxes represent classes, the rectangular ones represent the tags returned by POS tagger. Their super classes represent their grammatical functions (WH, verb, subject, etc.) and their subclasses represent the NL terms used in the question (When, Where, etc.). The solid edges represent the relation "subClassOf" and the dotted lines represent object properties. NL concepts are linked to their types through the "has_type" property and to their patterns through the "has_pattern" property. These elements are themselves subclasses of the concepts Types and Patterns respectively.

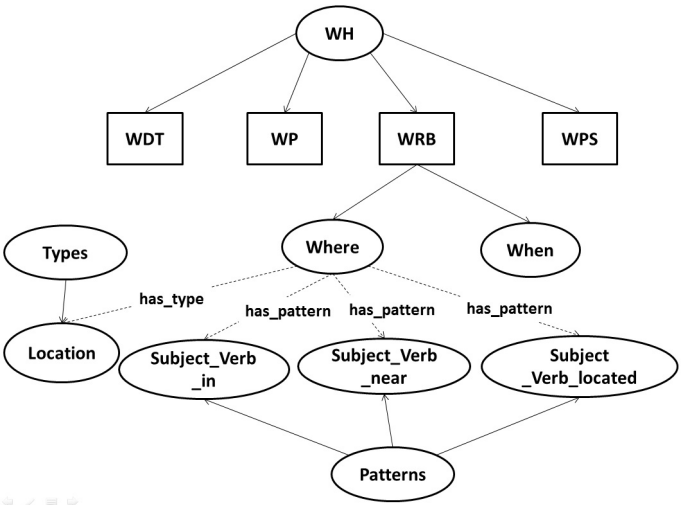

Fig. 3. Part of the question ontology

2. Instantiation of the generic graph: Using the parsed question, we instantiate the part of the generic question graph that contains the structure determined in the previous step. The result is a question graph that has the determined structure and contains the question's information. In fact, this graph is an instantiation of the generic graph that contains filled nodes of type "Term" containing the question's words. The terms of the same category form a graph and belong to the same type node: "WH", "Verb" or "Subject" (according to the example shown in Figure 2).Edges between these nodes are of type "followed_by" which specify the order of words in the user's question.
Example: Figure 4 is a question graph applied to the question "where is the tallest monument in the world?". This graph is an instantiation of the generic question graph shown in figure 2.

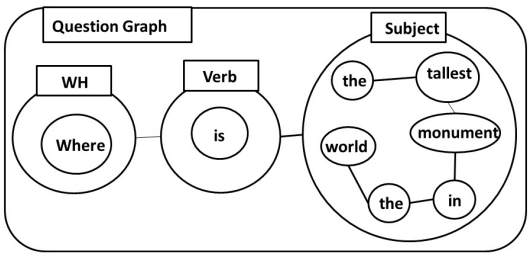

Fig. 4. Example of question graph

3. Synonym detection: WordNet is used in this step to extract the terms' semantics. We complete the question graph with the terms' synonyms in order to create a complete graph which contains the words, their grammatical functions, the structure of the question and synonyms. Adding synonyms to the graph is crucial for future search in the question base. In fact, the same question can be asked in several ways or expressed with different words and have the same meaning, in which case the system must be able to identify the different forms using the various synonyms stored in the question graph.

3.4 Query Reformulation Component

The analysis process requires query reformulation which consists on adding terms related to the question's keywords and expanding it. The resulting reformulated query will be submitted to the search engine that will return a set of documents from which the answer is extracted. The query reformulation is based, in our method, on two techniques which are the use of lexico-syntactic patterns and of a domain ontology. The aim is to guide the search engine to relevant documents for the search topic (using a domain ontology) and to answer-bearing documents (using patterns that define the answer's structure).

Query Reformulation Based on Patterns. The patterns used in this method are intended to reconstruct the user's question in order to guide it to the answer. Therefore, these answer patterns are applied to extract the candidate passage and locate the correct answer.

For each question type (what, where, who) there is an associated set of answer patterns. According to the question type of the submitted query, answer patterns are retrieved from the question ontology and instantiated with question terms. For instance, for the question: "where is the tallest monument in the world?", the method identifies from the question ontology the following patterns: Subject_Verb_in, Subject_Verb_near et Subject_Verb_located. The method reformulates the query using these patterns and obtains the following questions: "the tallest monument in the world is in", "the tallest monument in the world is near", "the tallest monument in the world is located".

Query Reformulation Based on a Domain Ontology. In order to add more semantic information to guide the search towards relevant documents, we use a domain ontology from which the method extracts, for the terms of the query that correspond to an ontology concept, its sub-classes and its related concepts. The method specializes the query by adding more specific concepts extracted from the ontology. This refinement increases the number of specific concepts and subsequently, increases the precision.

Let's take for example the question "where is the tallest monument in the world?" After the reformulation based on patterns in the previous section, we enrich the reformulated query with concepts related to the concept "monument" extracted from a domain ontology. We use the subclasses of this concept ("statue",

"arch ", "memorial") to enrich the reformulated query. We obtain three final reformulated queries: "the tallest monument/statue/arch/memorial in the world is in", "the tallest monument/statue/arch/memorial in the world is near" and "the tallest monument/statue/arch/memorial in the world is located".

3.5 Search for Similar Questions

This module of QAnalOnto retrieves similar questions stored in the question base. The method lists the similar questions ordered by similarity to the one asked by the user and, if the user chooses one, the corresponding analysis result and the reformulated query will be returned. However, the question base can be large, and the direct application of similarity measures can slow down the search process. To overcome this problem, we apply a filtering process that selects candidate questions from the base and removes questions that have total dissimilarity with the one asked by the user. In fact, the chosen questions from the base have the same expected answer type and at least one common keyword. On these questions, we will apply the similarity measures in order to classify them by their relevance to the user's question.

Several measures of semantic similarity, with different properties and results exist in the literature. The similarity measure we propose is based on the work of [11]. It combines the statistic similarity and the semantic similarity between the user's question and the questions stored in the question base. The statistic similarity is based on dynamically formed vectors: the two compared questions are represented with two vectors formed by their words instead of considering all the words in the question base and then their cosine product is computed to obtain the statistic similarity. The semantic similarity is calculated using the distance between two words w1 and w2 in WordNet as follows:

$$\frac{minDistToCommonParent}{DistFromCommonParentToRoot + minDistToCommonParent}$$

In this formula $minDistToCommonParent$ indicates the shortest path between two words to the common parent and $DistFromCommonParentToRoot$ indicates the path length from the common parent to the root.

The overall similarity is an average of statistical and semantic similarities.

4 Experimental Evaluation

A prototype has been developed to show that the proposed method can improve the performance of the retrieval task. It provides a user interface that allows these main functionalities: search for similar questions from the question base, construction of the question graph and reformulation of the user's query. Since the proposed method provides an analysis of the question and reformulates the query to be submitted to the search engine, we experimentally evaluate its performance by testing its capacity for (1) retrieving relevant documents after query reformulation and (2) retrieving similar questions from the question base.

4.1 Search Results Evaluation

To evaluate the query reformulation component, we computed: (1) Exact precision measures P@10, P@30, P@50 and P@100 representing respectively, the mean precision values at the top 10, 30, 50 and 100 returned documents; (2) MAP representing the Mean Average Precision computed over all topics.

Two main scenarios have been tested:

- The first scenario represents the baseline which is a classic search using keywords without performing any query reformulation.
- The second scenario represents results obtained after reformulating using both lexico-syntactic patterns and ontologies.

The improvement value is computed as follows:

$$Improvement = \frac{\text{Reformulation-result} - \text{Baseline-result}}{\text{Baseline-result}}$$

Table 1. Improvement in average precision at top n documents and MAP

	P@10	P@20	P@30	P@50	P@100	MAP
Baseline	0.60	0.32	0.212	0,171	0.065	0,273
QAnalOnto	0.783	0.39	0.256	0,206	0.078	0,341
Improvement	30,5%	21,87%	20.75%	20,46%	20%	24,90%

The evaluation results are calculated using the LEMUR [1] tool for Information Retrieval evaluation. Besides, we rely on the INEX 2010 [2] collection of documents. We measured the precision for several queries using the INEX topics and then we averaged these results. The evaluation results shown in table 1 represent the precision obtained according to the number of retrieved documents (10, 20, 30, 50 and 100), and we observe a significant improvement of the relevance of the retrieved information. In Table 1, we outline the computed MAP and the average precision at the top n documents and their percentages of improvement. We observe that reformulating queries using both lexico-syntactic patterns and a domain ontology improves the retrieval precision by 24,9%. In fact, using lexico-syntactic patterns guides the search towards answer-bearing documents and specifying the question's keywords and enriching it using the domain ontology improves the precision.

4.2 Similar Question Search Evaluation

To evaluate the search for similar questions, we used a set of queries (20 WH-questions from different domains). For each of them we created manually:(1) a set of questions containing the same words with different meanings, (2) a set

[1] `http://www.lemurproject.org/`
[2] `https://inex.mmci.uni-saarland.de/about.html`

of questions with different words but with the same structure and answer type and (3) one question with different words and the same meaning. In fact, this question is the only one considered similar to the tested question.

This set of questions is inserted into the question base. During the experimentations, we calculate the similarities between the user's question and each question extracted from the question base after filtering. We extract the most similar questions to the user's question and we return an ordered set of questions. To evaluate our method, the statistic, semantic and overall similarities have been calculated. For performance evaluation, we use the measures:

- Success at n (S@n), which means the percentage of queries for which we return the correct similar question in the top n (1, 2, 5, and 10) returned results. For example, s@1=50% means that the correct answer is at rank 1 for 50% of the queries.
- Mean Reciprocal Rank (MRR) calculated over all tested questions. The reciprocal rank is 1 divided by the rank of the similar question. The MRR is the average of the reciprocal ranks of results for the tested questions.

Table 2. s@n and MRR

	s@1	s@2	s@5	s@10	MRR
Semantic Similarity	15%	30%	60%	70%	0,338
Statistic Similarity	40%	70%	85%	95%	0,604
Overall Similarity	55%	95%	100%	100%	0,76

Table 2 represents s@n and the MRR measures that consider the rank of the correct similar question. The experimental results show that the overall similarity gives the best results and achieves a good performance. In fact, s@2=95%, that is to say for 95% of the questions, the similar question is extracted in 55% of the cases in the first position and 40% of the cases in the second.

5 Conclusion

This paper presents a new question analysis method based on ontologies. Our contribution can be summarized in: (1) representing the questions' structures by a generic graph; (2) representing the question by a typed attributed graph to ensure the representation of knowledge based on different levels of granularity; and (3) using lexico-syntactic patterns and domain ontologies to improve the query reformulation process and guide the search towards relevant (using domain ontologies) and answer-bearing documents (using patterns that define the structure of the answer).

Experiments were conducted and showed an improvement of the precision of information returned after the query reformulation and good similar questions extraction results.

As perspectives, we plan to develop automatic learning techniques to update the generic question graph and complete this work by adding an answer extraction method to search for answers in documents automatically.

Acknowledgments. This work has been supported by the Spanish-Tunisian AECID project A/030058/10, A Frameworkfor the Integration of Ontology Learning and Semantic Search.

References

1. Liu, H., Lin, X., Liu, C.: Research and Implementation of Ontological QA System based on FAQ. Journal of Convergence Information Technology Vol. 5, N. 3 (2010)
2. Hammo, B., Abu-salem, H., Lytinen, S., Evens, M.: QARAB: A Question Answering System to Support the Arabic Language. In: Workshop on Computational Approaches to Semitic Languages, ACL (2002)
3. Monceaux, L., Robba, I.: Les analyseurs syntaxiques: atouts pour une analyse des questions dans un systme de question-rponse. Actes de Traitement Automatique des Langues Naturelles, Nancy (2002)
4. Gaizauskas, R., Greenwood, M.A., Hepple, M., Roberts, I., Saggion, H.: The University of Sheffields TREC 2004 Q&A Experiments. In: Proceedings of the 13th Text REtrieval Conference (2004)
5. Wang, Y., Wang, W., Huang, C., Chang, T., Yen, Y.: Semantic Representation and Ontology Construction in the Question Answering System. In: CIT 2007, pp. 241–246 (2007)
6. Nyberg, E., Mitamura, T., Carbonell, J., Callan, J., Collins-Thompson, K., Czuba, K., Duggan, M., Hiyakumoto, L., Hu, N., Huang, Y., Ko, J., Lita, L., Murtagh, S., Pedro, V., Svoboda, D.: The Javelin question answering system at TREC 2002. In: Proceedings of the 11th Text Retrieval Conference (2002)
7. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
8. Tablan, V., Damljanovic, D., Bontcheva, K.: A Natural Language Query Interface to Structured Information. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 361–375. Springer, Heidelberg (2008)
9. Lopez, V., Uren, V., Motta, E., Pasin, M.: AquaLog: An ontology-driven question answering system for organizational semantic intranets (2007)
10. Kaufmann, E., Bernstein, A., Zumstein, R.: Querix: A natural language interface to query ontologies based on clarification dialogs. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 980–981. Springer, Heidelberg (2006)
11. Song, W., Feng, M., Gu, N., Wenyin, L.: Question Similarity Calculation for FAQ Answering. In: Proceedings of the Third International Conference on Semantics, Knowledge and Grid, pp. 298–301 (2007)
12. Moussa, A.M., Abdel-Kader, R.F.: QASYO: A Question Answering System for YAGO Ontology. International Journal of Database Theory and Application 4(2), 99–112 (2011)