Knowledge Base as an important element of the question-answering system

Tatiana PRODAN, Victoria LAZU
Technical University of Moldova
tatiana.prodann@gmail.com, lazu_vic@yahoo.com

***Abstract* – This paper describes the theoretical overview of the question-answering systems and knowledge base and representation. Within a joint project with the Information Society Development Institute (ISDI), a question-answering system for Romanian speakers is being developed. The document collection that was provided by ISDI serves as basis for the extraction of questions. The questions are thereafter classified according to the question types and thoroughly analyzed in order to generate the rules for the informal formalism of the knowledge representation process, enabling the system to retrieve the required information.**

***Index Terms* – knowledge, knowledge base, knowledge representation, question-answering system, question types**

I. INTRODUCTION

The creation of a question-answering system represents a complex process that involves many important components and resources. To this end, a project was launched, in cooperation with the Information Society Development Institute that aims to research the problems of Information Retrieval (IR) and create an automatic Question-Answering system for Romanian speakers. The system will process the queries in accordance with the Romanian language grammar and will retrieve the answers from the relevant texts. The first stage of project implementation foresees the development of a closed-domain question-answering system based on a legal texts collection.

II. QUESTION-ANSWERING SYSTEM

1.1. Brief history of question-answering systems

The question-answering (QA) systems are characterized by the fact that they receive a set of questions in natural language and have to extract answers to these questions from a collection of texts. This collection may differ from a simple local collection to the entire World Wide Web.

First question-answering systems have been developed in 1960 and were, in fact, interfaces that used natural language to interrogate expert systems created for various areas, among which the following can be mentioned [1]:

a) BASEBALL (Green, 1963) - answers to questions related to scores, teams, dates of the baseball games;
b) LUNAR (Woods, 1977) - accesses data of chemical field about moon rocks found during the Apollo missions;
c) PHLIQA1 (Scha, 1980) - answers to short questions about the data stored in a database containing a succession of specific information of the Phillips company.

A main feature of these systems is that the information was stored in a database that had to be developed by experts in the field.

SHRDLU [2] and ELIZA [3] systems followed after that. SHRDLU was a system able to answer questions about a world of objects, built from various geometrical forms that were moved by the system. ELIZA was able to simulate a conversation with a psychiatrist. Both were closed-domain question-answering systems.

During 1970-1980, several systems, which could interact with users applying natural language, have been developed. The most important are:

a) Unix Consultant - answered to questions about the UNIX operating system;
b) LILOG - was able to provide tourist information about a German city.

The flourishing period of the question-answering systems took place in the late 1990s when the Text Retrieval Conference (TREC) [4] included a section on QA systems. Modern question-answering systems, developed under the TREC impulse, are open-domain oriented systems. This requires a larger information base compared to the systems developed in 1960-1980.

Some interesting realizations of question-answering systems are also START [5] and AskJeevs [6] projects.

1.2. Methods used to develop question-answering systems

To create a question-answering system, two options are considered [7]:

a) Shallow approach - this method uses keywords to find passages and sentences in text that could stand as valid answers to questions. These potential answers are thoroughly analyzed to establish if the answers are real or not. This method can be successfully used in case of short and factual questions, when searching for names, dates, locations, quantities.
b) Deep approach - this involves a more sophisticated analysis, a syntactic, semantic and contextual processing. There are several methods that fall in this category: abduction, named-entity recognition, relation detection, etc.

Choosing one of the two methods depends on the complexity of questions that will be formulated and on the desired degree of system performance. The systems that use the deep method are superior to the ones using shallow method.

1.3. General architecture for a question-answering system

If at the beginning of the artificial intelligence, in 1960, the researchers were fascinated by the idea of creating systems able to answer questions belonging to restricted areas (closed domains), at the moment, the Internet development and the steps taken with regard to information retrieval and natural language processing techniques, as well as the demand for easy access to information, led to the increase of the interest for systems that would provide answers to open domain questions.

A question-answering system based on a collection of documents typically includes three main components:

a) Question analysis module - that transforms the questions posed in natural language to queries for the document retrieval engine;
b) Document retrieval module - that searches in the collection of documents the relevant articles for the question submitted by the user, based on data received from the question analysis module;
c) Answer extraction module - from the collection of articles returned by the document retrieval module, it extracts a concise answer that is also a natural language response to the user's question. If such an answer does not exist in the collection of documents considered by the document retrieval module, it is preferred that the system does not answer the question, instead of returning a wrong answer.

The general construction for a question-answering system could be conceptually represented as shown in Figure 2.1.

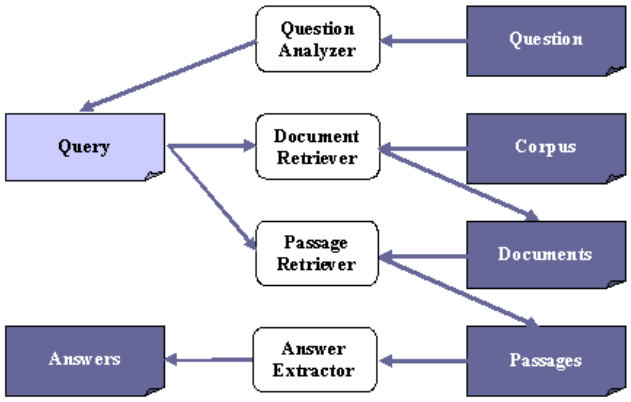

Fig. 2.1. General architecture for a question-answering system

1.4. Problems in creating a question-answering system

Although Internet is an environment full of information in all areas, finding an answer to a simple question can sometimes be a difficult task. Some problems that may arise in developing a question-answering system are:

a) Correct formulation of queries - transforming a natural language question to a query for a search engine is a tall order. If the question is too general, a large number of documents will be extracted. The topics described in the retrieved collection of documents may not contain the exact answer to the user's question. If the set of words is too small, the article containing the answer to the query may not be found. Therefore, it is required to formulate well the question so that the system retrieves documents containing useful information.
b) Retrieval of correct answers - even if the appropriate set of words is found to make a query that would retrieve useful information, the search engine may retrieve many articles that do not answer the user's question.
c) False information - even if the question is well formulated and the search engine retrieves articles that correspond to the topic of the query, it may be possible that some of these articles contain wrong data.
d) Limited resources - when a question-answering system is created, it is necessary to take account of the restrictions imposed by the processing of large amounts of information. It is not advisable to send to the system a long set of word strings. Searching in a long list of articles is time-consuming and the user is not willing to wait too much for an answer.

III. KNOWLEDGE BASE

An essential element of a question-answering system is the knowledge base (KB) [8]. In order to develop such a system it is required to capture and express knowledge in the form of rules. The knowledge base includes the taxonomy of concepts and relations and the representation of their interconnections.

The ability of the question-answering system to provide appropriate results depends on the quality and volume of knowledge available to it. Typically, the data flow of a question-answering system takes place as shown in Figure 3.1.

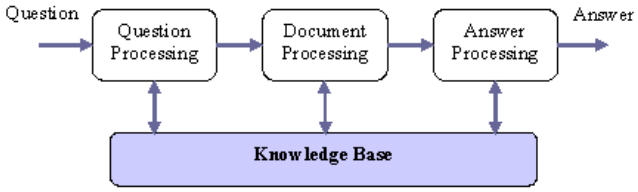

Fig. 3.1. Data flow in a question-answering system

As seen in the figure above, the knowledge base is critical for each operation that takes place in the question-answering system.

The development of a QA system involves tasks such as obtaining of knowledge, its documentation and organization, generation of a knowledge network that exemplifies the relationship between different sources of knowledge, verification of knowledge consistency and transformation of the knowledge network in a computer program using specific instruments. Such a programme is a formal system for storage of facts and relations between them and of the strategies for their use. In addition, different search mechanisms have to be represented in order to operate the system.

Generally, the knowledge of a KB is grounded on three fundamental concepts:

a) facts, representing the primary information that describes the elements of the considered domain;
b) rules, describing how to use facts;
c) reasoning strategies, expressing the manner in which rules can be used.

It is worth-mentioning that while creating our question-answering system, we aim at enabling the knowledge base to continuously grow without affecting its smooth running and without making any substantial changes to it.

In order to store and use knowledge the knowledge structures are applied, just as to store and process data the data structures are used.

The storing of knowledge in a computer-readable form consists in finding a correspondence between the outside world and the symbolic system that allows the execution of reasoning. For this purpose, those characteristics that can be associated with the meanings determined in the mental image about the world are extracted from the observations made over the objects, facts and phenomena.

The description aims to distinguish the image of the given object from the image of other surrounding objects. This shows that the description reflects the knowledge about the object itself. The knowledge is related to a "knowledgeable subject" that has the capacity to interpret the information that is available to him by extracting it from the own memory. Like this, the knowledge has an individual character, as people can possess different levels of knowledge depending on the represented things. Therefore, knowledge is usually partial and/or incomplete. On the other hand, knowledge can have an empirical character.

Taking account of these characteristics, the major problem of artificial intelligence is to define methods for representing large amounts of knowledge in a form that allows its efficient storage and usage.

To this end, it is necessary to point out that the formalization of knowledge and of strategies for answering questions form a major part in creation of a QA system. The same knowledge can be represented using a formal scheme, but with difficulty variations. The difficulty does not consist in knowledge representation but in its use. As well, different knowledge representation schemes can be adapted to develop an application.

In order to enable a computer to process the query and provide an appropriate answer a well-defined description of the problem is required. Like this, it is necessary to express the question in natural language and then represent it in formal language so that computer can interpret it. After that, the computer uses an algorithm to compute the corresponding answer. The Figure 3.2 illustrates this process.

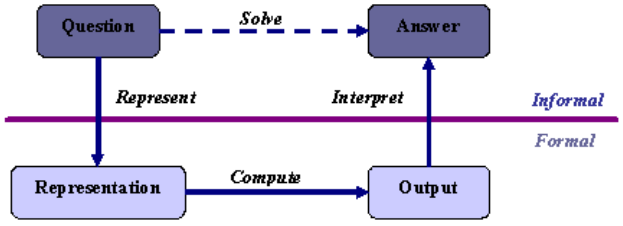

Fig. 3.2. Knowledge representation scheme

The steps that are taken in the above scheme are as follows:

a) first, the informal formalism of the question is carried out;
b) then, it is formally represented and the computer provides the output;
c) the output is expressed in an informal answer that is understood by the user.

The most common methods of knowledge representation [9,10] are:

a) predicate logic
b) production rules
c) semantic networks and frames

The **predicate logic** provides mechanisms to represent the facts and reasoning based on syntactic handling of logical formulas. This method uses predefined inference rules to declare and deduct facts. In the first-order predicate logic, the formulas are handled according to their form or structure. The great disadvantage of this scheme is that the meaning or semantic context of the formulas cannot be considered.

In predicate logic, all inferences are based on logical deduction and the inference rules are correct. In addition, a logical program will generate all possible inferences, taking into account all facts and rules. The systems based on predicate logic try to apply all inference rules to all facts. There is no mechanism for grouping the facts and associating the specific inference rules to different groups.

The **production rules** are simple forms for representing knowledge, which provide the flexibility to combine declarative and procedural representations in order to use them in a unified form. A production rule has a lot of assumptions and conclusions. The hypothesis specifies a set of conditions and consequences of certain actions. Typically, a knowledge base will consist of a multitude of rules. Logically, the rules can be grouped into different rule bases. A knowledge network representing a multitude of rules should be complemented with appropriate connections within the network nodes. Hence, the drawing of the knowledge network provides the opportunity to check the knowledge base for possible inconsistencies and redundancies.

A **knowledge network** consists of a set of nodes and arcs. The nodes are used to represent the concepts or entities and the arcs represent the relations between the concepts. The search for new properties in the hierarchy of concepts is assimilated with an inference by inheritance. To obtain a higher degree of generalization, structural relations can be introduced to express the intrinsic properties of concepts associated to nodes. To this end, there are relations for allocation of generic properties that link a concept to its attributes and relations for assigning specific properties that link a concept representative to its specific properties.

Frames refer to a representation method that allows the combination, in a unique structure, both of declarative and procedural knowledge. Frames belong to inheritance transfer methods. Along with the knowledge storing function they organize the knowledge in hierarchies or inheritance networks through which the sharing of common properties takes place. Each element of the frames hierarchy inherits the properties of the elements located on the upper level. By connecting the frames with a set of production rules, an integrated system for representation and inference is obtained.

IV. CURRENT IMPLEMENTATION AND RESULTS

The question-answering system that started to be developed within the above-mentioned project is based on

legal framework documents of the Republic of Moldova and is intended to Romanian speakers. The system aims to facilitate the quick identification of a law that is required to be applied in a certain context and provide appropriate answers to users' questions. At the same time, the system will be available for anyone.

At the first stage of implementation a collection of documents was provided to us by the Information Society Development Institute, including:

a) Code on science and innovation of the Republic of Moldova;
b) Decision on approval of the Partnership Agreement between the Government and Academy of Sciences of Moldova for the years 2009-2012;
c) Law on scientific and technological parks and innovation incubators.

These documents have been thoroughly analyzed and related questions have been manually created. As for instance, for the "Code on science and innovation of the Republic of Moldova" 134 questions were generated.

The next activity was focused on classification of the created questions according to the QA@CLEF [11] question categories. Like this, there were defined some major types of the extracted questions, among which:

a) **Factoid**: Ce reglementează codul cu privire la ştiinţă şi inovare?
b) **Definition**: Ce este dezvoltare tehnologică?
c) **List**: Care sunt atribuţiile Guvernului în sfera ştiinţei şi inovării?
d) **Explanation**: Cu ce se ocupă Agenţia de Stat pentru Proprietatea Intelectuală?
e) **Count**: De câte ori pe an se convoacă Comisia de acreditare a organizaţiilor din sfera ştiinţei şi inovării?
f) **Time**: În ce an a fost adoptat Codul cu privire la ştiinţă şi inovare al Republicii Moldova?

Nonetheless, there is a large diversification of question types, which can be more general or more detailed. The number of categories and their properties are defined according to the goal of the developed application and its capacities.

The following step is the creation of rules for conversion of the natural language questions into the formal semantic queries. Like this the informal formalism takes place. Some examples of simple rules are:

a) <ACT> reglementează <ACŢIUNE> or <ACŢIUNE> este reglementată <ACT>
b) <OBIECT/NUME> este <DEFINIŢIE>
c) <ATRIBUTII> sunt <LISTA>
d) IF <când/cînd> THEN <timp>
e) IF <cine> THEN <nume>

For a proper formalization of the natural language queries, there is also intended to apply semantic graphs and frames.

As well, some problems have been detected to this end, including the following:

a) Lack of the keyword in the answer: for example "Comisia: …" - from the context it is easy to understand that there are listed the attributions of the Commission, although the keyword "attributions" is not present in the answer and for the system this creates the impossibility to retrieve the answer to the user's query.
b) Lack of the predicate in many definitions: for example "Agenţia – o instituţie …", like this the verb is replaced with a hyphen.
c) The full name of a legal body mentioned in the available documents is written only at the beginning of the chapter that regulates its activity, and further only the word "Agency" or "Commission" is present: for example "Consiliul Naţional pentru Acreditare şi Atestare" is mentioned entirely only once at the beginning of the law chapter, although in further articles of the chapter it appears as "Consiliu" and when a query is entered, the system may not be able to provide an appropriate answer.

V. CONCLUSION

This paper tackles the issue of knowledge base that is a critical component of the question-answering system. The first steps implemented so far to develop the system were the analysis of document collection, the extraction of questions and their classification and the creation of rules.

REFERENCES

[1] I. Androutsopoulos, G. D. Ritchie, P. Thanisch, "Natural Language Interfaces to Databases - An Introduction", Natural Language Engineering, 1(1), p. 29-81, 1995.
[2] T. Winograd, "Understanding Natural Language". Academic Press, New York, 1972.
[3] J. Weizenbaum, "ELIZA - a computer program for the study of natural language communication between man and machine", Communications of the ACM 9 (1), p. 36–45, January 1966.
[4] Information on Text Retrieval Conference available on: http://trec.nist.gov/
[5] Information on START question-answering system available on: http://start.csail.mit.edu/
[6] Information on AskJeevs question-answering system available on: http://www.ask.com/
[7] M. T. http://en.wikipedia.org/wiki/Question_answering#cite_ref-4Maybury, "New Directions in Question Answering", 2004. AAAI/MIT Press.
[8] C. Periñán-Pascual, F. Arcas-Túnez , "Cognitive Modules of an NLP Knowledge Base for Language Understanding", Natural Language processing, No. 39, 2007, pp.197-204
[9] R. C. Chakraborty, Course on Artificial Intelligence, Department of Computer Science and Engineering, Jaypee University of Engineering and Technology, Guna.
[10] M. Bosch, "Ontologies, Different Reasoning Strategies, Different Logics, Different Kinds of Knowledge Representation: Working Together". Knowledge Organization, Ergon Verlag. 33(3) 153-159, 2006. ISSN 0943-7444 Knowl.Org.
[11] D. Giampiccolo, P. Forner, A. Peñas, C. Ayache, D. Cristea, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu and R. Sutcliffe, "Overview of the CLEF 2007 Multilingual Question Answering Track". Online proceedings of CLEF 2007 Working Notes, Budapest, September, 2007, ISBN: 2-912335-31-0.