International Journal of Advances in Computer Science & Its Applications – IJCSIA
Volume 4: Issue 3 [ISSN 2250-3765]
Publication Date : 30 September,2014

Question Answering Systems in E-Government: A Proposal for the Colombian Context

[Adán Beltrán Gomez, Leonardo Melo Gonzalez]

***Abstract*—Given the staggering quantity of textual information on the Internet, several systems of information search and retrieval have been proposed in recent years, such as Question-Answering Systems (QAS), whereby users employ natural language in their search, retrieving a specific answer to their question instead of a list of relevant documents. Government websites publish a wide array of textual information such as laws, procedures, and policies. Users generally do not perform searches in government websites due to the complexity of search mechanisms, which are sometimes too technical for lay people. The article proposes an architecture for the integration of QAS with government web portals in order to improve the search process on the part of citizens, thereby improving the interaction between citizens and government. It also discusses the research findings related to the design and implementation of a QAS in the Colombian context proposed for a local government.**

***Keywords*—e-government, Question Answering, citizens interaction**

I. Introduction

Due to the large amount of textual information currently available on the Internet, Web users spend a considerable amount of time trying to find relevant and accurate information, despite the powerful search engines available. Additionally, search functionality on many websites does not work properly, prompting users to "navigate" between different links and read the contents of the files, thereby degrading usability and experience; users end up abandoning the websites and looking for other mechanisms to find information, such as phone calls to the companies´ call centers or offices.

Different disciplines have proposed mechanisms to improve the search for relevant information by the user, such as Information Retrieval (IR), which has provided the theoretical basis of current search engines; Information Extraction (IE), based on systems that extract knowledge from texts; and the semantic Web, which has produced an architecture based on standards that facilitate automatic interaction between applications through semantic labeling of data. Among the Information Retrieval systems are Question-Answering Systems (QAS), which aim to give the user a specific answer to a question instead of a list of relevant documents. These systems, that allow the user to perform natural language queries, generally have a structure of algorithmic complexity that involve the use of techniques of Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP), among others. In general, the complexity of these systems lies in establishing the implied relationship between a question and an answer. Certain approaches consider them as systems on the way to the Semantic web (Konopík & Rohlík, 2010).

In this article the authors discuss the results of research in the development of a QAS for the Colombian Government online. In the first part the general concepts of QAS and the status of e-government in Colombia are introduced, and in the second part the results of a research in the design of a domain-specific QAS for interaction between citizens and the Colombian state are discussed.

Adán Beltrán Gómez
Universidad Manuela Beltrán
Colombia

Leonardo Melo Gutierrez
Pontificia Universidad Javeriana
Colombia

II. Background

A. *Question Answering*

Even though QAS have evolved considerably since its inception in the 60s with systems like Eliza (Weizenbaum, 1966), Baseball (Green Jr, Wolf, Chomsky, & Laughery, 1961) and Lunar (Woods, 1973), among others, their precision in response generation is still low and it is not the same for all languages: in English, for example, over 77% of the questions are answered correctly, according to an analysis by Voorhees (Voorhees, 2004) based on data from the TREC-2004 Conference, but in Spanish this percentage falls to 55% according to analysis of data from the CLEF 2006 conference (Bernardo Magnini et al., 2007), even though Spanish is the third language with more presence in the web and the second language used for online queries. This might be due in part to the few corpora available for this language.

There are basically two approaches for the development of QAS: semantic and statistical. From the semantic point of view, Sowa (Sowa, 2011) proposes three levels of the semantic knowledge that can be obtained: Strong, where

information is labeled using formal logic, allowing for a large reasoning capacity, Medium, in which the information has a semi-formal notation, allowing for a certain level of reasoning, and Light, that uses classification and restriction levels but no reasoning is possible. Systems may be of one of these types, or can combine some of these levels in their components.

Given the potential applications of QAS, in recent decades different proposals have been developed, both in open domain systems (B Magnini, 2005) and in specific domain systems: education (Wang, 2012; Wen, Cuzzola, Brown, & Kinshuk, 2012), virtual tutors (Olney, Graesser, & Person, 2010; Tewari, Liu, Cai, & Canny, 2012), legal (Hughes, Hughes, & Lazar, 2008), tourism (Todorova & Gelfond, 2012), treatment of diseases (Pechsiri, Painuall, & Janviriyasopak, 2012), access to data warehouses (Kuchmann-BEAUGER & Aufaure, 2011), medical (Lu, Tung, & Lin, 2010; Yong-Gang, Ely, & Hong, 2009), among others. Despite the progress, sophistication and potential applications of these systems, such as the IBM Watson system, they are not entirely accurate and there is still much to improve.

B. E-Government in the world and in Colombia

The definitions of e-government are varied, have changed over time, and are not identical in all authors. The definitions of various authors mentioned by Yildiz (2007), for example, emphasize different aspects. Most definitions, however, conceive e-government as the use of information and communication technologies (ICTs) for a more efficient management of public sector functions, and better communication between government and citizens.

The implementation of e-government in four stages proposed by Layne and Lee (2001) has been the pattern followed by most governments in the world. Based on the experiences of various government agencies, the authors propose a four-phase implementation that takes into account the technical, organizational and administrative limitations of different organizations: cataloging, transaction, vertical integration and horizontal integration. These phases involve increasing levels of technological and organizational complexity, and higher levels of integration of services and functionalities. In the first stage, cataloging, governments are focused on cataloging government information and displaying it on the Web. This stage completed, citizens will find in government sites information, forms and indexes, with little or no interaction from users. In the next stage, transaction, sites allow citizens to perform online transactions such as payment of fines and license renewals that involve interacting with government databases. The third stage, vertical integration, based on the experience with American levels of government, connect services and information from local, state and federal governments with similar functions, and the fourth stage, horizontal integration, refers to the integration of different services and functions across all government agencies, so that citizens can perform their searches and transactions from a single portal. This last stage will consolidate the communication between government and citizens (G2C), government and business (G2B), and between government agencies (G2G).

The United Nations has developed an index of development of e-government that allows comparisons between countries and across time, which gives an idea of "the willingness and ability of national governments to use information and communications technology in the delivery of public services" (United Nations, 2012). This index is a composite indicator, the average of three indexes rated from 0 to 1 measuring three essential dimensions of e-government: scope and quality of services online, telecommunications infrastructure development, and human capital. Of the 193 member states reported in the *United Nations E-Government Survey 2012*, the first three places correspond to the Republic of Korea (0.9283), the Netherlands (0.9125) and the UK (0.8960). At the other end of the spectrum, excepting the three countries that do not have an e-government initiative (Central African Republic, Guinea and Libya), we find in the bottom three places Niger (0.1119), Chad (0.1092), and Somalia (0.0640). In this spectrum Colombia, ranked 43 with a value of 0.6572, is located toward the bottom of the first quartile. Compared to its neighbors in the Americas, Colombia ranks fourth behind the United States (0.8687), Canada (0.8430) and Chile (0.6769), placing Colombia in second place in Latin America in terms of e-government development. All these indicators point to the advanced development of e-government in Colombia, the result of efforts of the national government to improve communication and interaction between citizens and government through the use of ICTs.

According to the first measurements done by the Colombian Government in 2009, the national agencies were more advanced than the territorial ones in implementing e-government; sectors of the state with more electronic presence were the executive, legislative and judicial branches, electoral organization, control agencies and autonomous bodies; 31% of citizens and 72% of large companies had used the Government Online platform; and the most popular online transaction was request for Judicial Certificate of the Administrative Department of Security (DAS)(Massal and Sandoval, 2010).

Finally, according to the Colombian report on e-government (*El gobierno en línea en Colombia 2011*), the Connectivity Agenda has implemented monitoring and evaluation systems that generate information on the progress of e-government and its perception by citizens and businesses. Broadly speaking, this evaluation supports the notion that citizens and businesses have gradually gained confidence in e-government and are using it increasingly, although there are still important differences between national and territorial agencies. To conclude this discussion, it is worth noting that while there has been significant progress in the components of

online services and human capital, which are relatively high, there is still much to do in infrastructure for Colombia to improve its implementation indicators of e-government (Republic of Colombia, 2011).

In this context, the research project proposes the development of a QAS in the context of e-government, a pilot test of the system in a local government and its appropriation by citizens.

III. Proposed System

One of the first tasks in the research process was to define the basic architecture of the system for subsequent design, development and implementation of components. Proposals for different architectures for QAS such as AQUALOG (Lopez, Uren, Motta, & Pasin, 2007), QALLME (Ferrández et al., 2011), FALCON (Harabagiu et al., 2000), among others, were reviewed. Such architectures have much in common, including the fact that they are based on integrated processes through input and output.

A. *Architecture*

Several information resources were considered as key elements for both the training of the system as well as a basis for the system to answer certain types of questions. Data sources that were taken into account were: FAQ files that state agencies publish as a quick guide for citizens, and texts retrieved from websites of different government sectors. In Fig. 1 the components of the Framework architecture that allow processing of questions and their best option solutions are described.

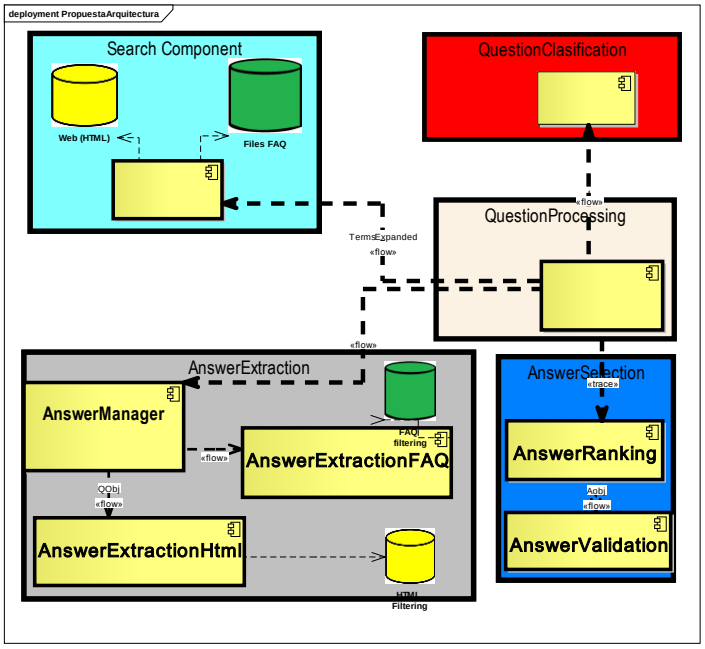

Figure 1. Architecture of proposed system

Following is a description of the most important components of the proposed architecture:

• Question Processing. It is the main component of the system and is responsible for managing and orchestrating system components. It takes as input a user query in natural language and relates these questions to the respective components.

• Question Classification. This component determines the keywords, the question type and expected response using techniques for automatic classification for the categories of the defined taxonomy for the specific domain. It expands the question with ontology terms and EuroWordNet, returning the expanded questions to the Component Search module.

• Search Component. It takes the pre-processed and expanded questions, searches in the database of FAQs saved HTML documents and web searches, returning the documents which could hold the responses to questions.

• Answer Extraction. Extracts the response from the filtered FAQ database and the paragraph where the answer could be, based on the type of response expected using the automatic text classifier of responses.

• Answer Selection. Because the system extracts possible answers from different sources of information, it has to select the best answer using ranking algorithms to extract the best response from the document, checks whether the selected response corresponds to the question entered initially, and returns the response.

B. *Resources*

A series of resources were collected and built that were the basis for the system components; the most important of them were:

• Expect Answer Type. It collects categories of questions and answers that can be answered. This taxonomy was defined from the analysis of the intentions of citizens of the collected sample and from the sectors into which the Colombian government has organized its services. A total of 11 categories were defined.

• Corpus of FAQ files. From the Government Online initiative was obtained the file containing information from all agencies of the state, from which the XML files were generated with organized questions. The collection of FAQs was obtained visiting the web pages of the 863 agencies of the Colombian state, which after removal of repeated questions allowed the compilation of 8241 FAQs from different categories.

• Ontology and RDF files. To increase the likelihood of success at the time of answering a question, the ontology that

associates terms according to a given context was defined, establishing groups of words that allow to extend the questions from the keywords.

IV. Results

For the precision tests of the system a corpus of questions and correct answers was built. The size of the corpus of questions was of 51, whereas the collections containing the answers was of 300, corresponding to the different categories of the proposed taxonomy (Fig. 2 and Fig 3).

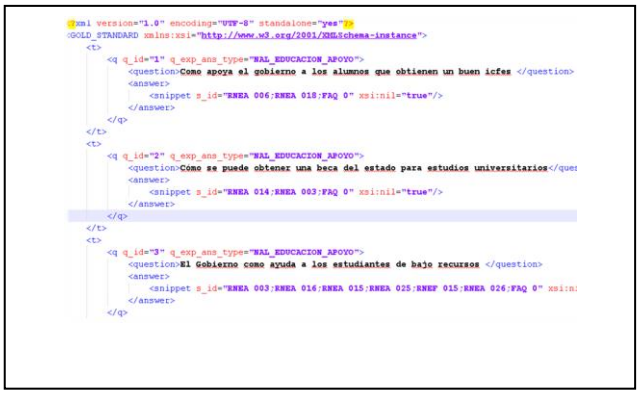

Figure 2. Example of questions and corresponding EAT (Expect Answer Type) category.

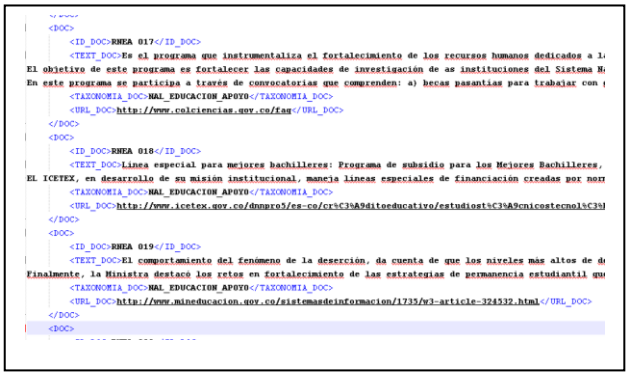

Figure 3. Example of answer documents collection

As is evidenced on Fig. 4, 71% of documents classified or selected by the system coincided with the expected ones.

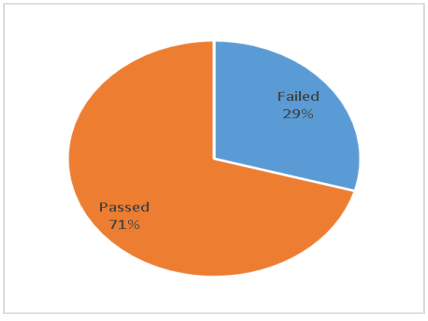

Figure 4. System precision

Regarding the tests with actual citizens, a screen aid was added (Fig. 5) to help citizens write the questions according to the FAQ collection loaded to the system. Citizens´ perceptions was evaluated during a week, to determine users´ experience: an average 80% said the system would help them to interact with government portals. The project hopes to continue evaluating perceptions of citizens regarding QAS in e-government.

Figure 5. Graphical interface of the system

V. Conclusions

This article has discussed the process of research and development of a QAS in the domain of e-government. Due to the large number of sectors of the government, the specificity is low because it is not possible to select only one sector, so that the term "domain specific" is debatable.

The proposed architecture for the development of a QAS has the following features compared to the systems considered:

• It combines different approaches: use of FAQ files, query expansion through ontologies and EuroWordNet, and use of Semantic Web elements.

• It integrates technologies such as web services developed in Python for natural language processing, web services Jena based on Java for the use of ontologies, and web client for the use of services.

• It uses natural language processing technologies to improve the usability of web sites, as users will not have to browse websites looking for information.

• A first sample shows 80 % of citizens think this technology is interesting and helps them to interact with state services.

• An important corpus to train and test these systems in the context of e-government in general for the Spanish language has been built.

Acknowledgment

Research Project funded by Colciencias, Ministry of Information Technologies and Communications of Colombia (MinTic), Manuela Beltrán University, Multimedia & Service ltd and IP Innova Ltda - Project No. 1263-595-37045.

References

[1] Ferrández, O., Spurk, C., Kouylekov, M., Dornescu, I., Ferrández, S., Negri, M., . . . Vicedo, J. L. (2011). The QALL-ME framework: A specifiable-domain multilingual question answering architecture. *Journal of Web Semantics, 9*(2), 137-145.

[2] Forner, P., Peñas, A., Agirre, E., Alegria, I., Forăscu, C., Moreau, N., . . . Tjong Kim Sang, E. (2009) Overview of the clef 2008 multilingual question answering track. *9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008: Vol. 5706 LNCS* (pp. 262-295).

[3] Green Jr, B. F., Wolf, A. K., Chomsky, C., & Laughery, K. (1961). *Baseball: an automatic question-answerer.* Paper presented at the Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference.

[4] Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., . . . Morarescu, P. (2000). *Falcon: Boosting knowledge for answer engines.*

[5] Hughes, T., Hughes, C., & Lazar, A. (2008). Epistemic structured representation for legal transcript analysis *Advances in Computer and Information Sciences and Engineering* (pp. 101-107): Springer.

[6] Konopík, M., & Rohlík, O. (2010) Question answering for not yet semantic web. *13th International Conference on Text, Speech and Dialogue, TSD 2010: Vol. 6231 LNAI* (pp. 125-132).

[7] Kuchmann-Beauger, N., & Aufaure, M. A. (2011) A natural language interface for data warehouse question answering. *16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011: Vol. 6716 LNCS* (pp. 201-208).

[8] Layne, K., Lee, J. (2001). Developing fully functional e-government: a four stage model. *Government Information Quarterly, 18* (2), 122-136.

[9] Lopez, V., Uren, V., Motta, E., & Pasin, M. (2007). AquaLog: An ontology-driven question answering system for organizational semantic intranets. *Web Semantics: Science, Services and Agents on the World Wide Web, 5*(2), 72-105.

[10] Lu, W.-H., Tung, C.-M., & Lin, C.-W. (2010). Question Intention Analysis and Entropy-Based Paragraph Extraction for Medical Question Answering. In C. T. Lim & J. C. H. Goh (Eds.), *6th World Congress of Biomechanics (WCB 2010). August 1-6, 2010 Singapore* (Vol. 31, pp. 1582-1586)

[11] Magnini, B. (2005). *Open Domain Question Answering: Techniques, Systems and Evaluation.* Paper presented at the Tutorial of the Conference on Recent Advances in Natural Language Processing-RANLP.

[12] Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., . . . Sutcliffe, R. (2007). Overview of the CLEF 2006 multilingual question answering track *Evaluation of Multilingual and Multi-modal Information Retrieval* (pp. 223-256).

[13] Massal, J., Sandoval, C.G. (2010). Gobierno electrónico ¿estado, ciudadanía y democracia en Internet? *Análisis Político, 23*(68), 3-25.

[14] Olney, A. M., Graesser, A. C., & Person, N. K. (2010). Tutorial Dialog in Natural Language.

[15] Pechsiri, C., Painuall, S., & Janviriyasopak, U. (2012). Medicinal Property Knowledge Extraction from Herbal Documents for Supporting Question Answering System. In L. Cao, J. Huang, J. Bailey, Y. Koh & J. Luo (Eds.), *New Frontiers in Applied Data Mining* (Vol. 7104, pp. 431-443).

[16] República de Colombia. (2011). *El Gobierno en línea en Colombia 2011.*

[17] Schmid, H. (1995). TreeTagger a Language Independent Part-of-speech Tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43*.

[18] Sekine, S., Sudo, K., & Nobata, C. (2002). *Extended Named Entity Hierarchy.*

[19] Sowa, J. F. (2011). Future Directions for Semantic Systems.

[20] Tewari, A., Liu, I., Cai, C., & Canny, J. (2012) An analysis of the dialogic complexities in designing a question/answering based conversational agent for preschoolers. *12th International Conference on Intelligent Virtual Agents, IVA 2012: Vol. 7502 LNAI* (pp. 36-45).

[21] Todorova, Y., & Gelfond, M. (2012) Toward question answering in travel domains. *Vol. 7265* (pp. 311-326).

[22] United Nations. (2012). *Estudio de las Naciones Unidas sobre el gobierno electrónico 2012.*

[23] Voorhees, E. M. (2004). Overview of the TREC 2004 question answering track (pp. 52–62).

[24] Wang, W. W. Z. L. S. (2012). Preference Survey and System Design on Mobile Q&A for Junior High School Students.

[25] Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

[26] Wen, D., Cuzzola, J., Brown, L., & Kinshuk. (2012) Exploiting semantic roles for asynchronous question answering in an educational setting. *25th Canadian Conference on Artificial Intelligence, AI 2012: Vol. 7310 LNAI* (pp. 374-379).

[27] Woods, W. A. (1973). *Progress in natural language understanding: An application to lunar geology.* Paper presented at the Proceedings of the June 4-8.

[28] Yildiz, M. (2007). E-government research: reviewing the literature, limitations, and ways forward. *Government Information Quarterly, 24*(3), 646-665.

[29] Yong-Gang, C., Ely, J., & Hong, Y. (2009, 1-4 Nov. 2009). *Evaluating the weighted-keyword model to improve clinical question answering.* Paper presented at the Bioinformatics and Biomedicine Workshop, 2009. BIBMW 2009.

[30] Zhang, D. (2012). Inconsistency in Multi-Agent Systems.

About Author (s):

Adán Beltrán Gómez, Master of Science in Information and Communication University Francisco José de Caldas, Bogotá, Colombia, Research project leader, Universidad Manuela Beltran,

.

Leonardo Melo Gonzalez, Master in Library and Information Science at the University of Texas at Austin, Professor at Pontificia Universidad Javeriana,