SEARCH FOR ANSWERS IN DOMAIN-SPECIFIC SUPPORTED BY INTELLIGENT AGENTS

Fernando Zacarias[1], Rosalba Cuapa [2], Guillermo De Ita[1] and Miguel Bracamontes[1]

[1]Computer Science, Universidad Autonoma de Puebla, Puebla, México.
[2]Faculty of Architecture, Universidad Autónoma de Puebla, Puebla, México.

ABSTRACT

Search for answers in specific domains is a new milestone in question answering. Traditionally, question answering has focused on general domain questions. Thus, the most relevant answers (or passages) are selected according to the type of question and the Named Entities included in the possible answers. In this paper, we present a novel approach on question answering over specific (or technical) domains. This proposal allows us to answer questions such as "What article is appropriate for … ", "What are the articles related to … ", these kind of questions cannot be answered by a general question answering system. Our approach is based on a set of laws of a specific domain, which contain a large set of laws regarding the work organized into a hierarchy. We consider generic concepts such as "article" semantic categories. Our results on the corpus of Federal Labor Law show that this approach is effective and highly reliable.

KEYWORDS

Question Answering, Search for Answers, Mobiles, Intelligent Agents, & Question Answering

1. INTRODUCTION

Question Answering (QA – for its acronym in English) is an area that has grown rapidly in recent years, however this has not been able to respond concretely to specific questions on technical or specific contexts, although this being the main goal of QA. Currently, many proposals are focused on answering specific questions in specialized context and have taken much more relevance. This is because some groups of professionals are reluctant to use general search engines in their professional activities because they do not receive satisfactory and concrete answers. Those professionals have their own work habits, and they consider browsing the Web as an inefficient way to find professional information. In order to satisfy these professionals must consider some aspects such as build machines search for specialized contexts and build search engines that provide accurate answers to your questions. This makes it possible to improve search engines with a specialized QA.

While a lot of effort has been spent on general-purpose QA, few research projects have focused on domain-specific QA, such as in the labor sector. One of the major problems in the area of labor expertise lies primarily in two broad themes: First, know each and every one of the existing laws within the legal framework of the question and second, identify which of these laws may be applicable to a particular problem into the labor sector. For this reason, we have developed various strategies to address these problems, one is seeking answers to specific questions within a particular area, as is the case of "The Labor Law".

The search for answers through mobile devices is an issue that is growing rapidly in recent years, due to growing demand from users to obtain satisfactory answers to their needs on large volumes of digital information currently available. Moreover, mobile search becomes more relevant for the usability of mobile content, similar to how our machines look online at the usability of web content. Currently there are not companies that do not offer a mobile service of some kind. Many

DOI: 10.5121/ijcsit.2018.10404

operators in internet portals show their best content available. However, much of the content developed for mobile devices pass unnoticed by users.

Accordingly, we developed a novel approach based on a fully automated algorithm and supervised by intelligent agents that allows through a mobile device look for answers that users demand about labor law. Thus, contrary to what internet systems perform, we seek for answers concrete to problems related with labor law into the labor sector (specifically for Mexico). Furthermore, the rapid emergence and growth of mobile technologies has detonated the development of mobile applications that allow users to satisfy many of their needs from anywhere, from any mobile device and any time. In addition, we take advantage of the fact that for every personal computer currently there are more than six mobile devices. In the United States in 2013, 91% of adults have a cell phone, and of these, 63% use it to connect to internet [1]. This represents an excellent opportunity for new proposals as presented in this paper. New trends in the area of question answering has emerged as a new interesting research domain, various proposals are still based predominantly on processing questions and methodologies for the extraction of answers. Today, question-answering systems are primarily focused on to answer questions from casual users, i.e., from the point of view of a user who does not have expertise in the field of treating search. Therefore, questions have as answer a fact, situation or specific data. Gradually has increased the level of difficulty of questions, for instance, questions whose answer is a list of instances or a definition. For this reason, in this paper we extend the existing approaches to QA systems to deal with domain-specific questions. In particular, our algorithm makes use of a set of laws that apply to a specific domain called "Federal Labor Law", which governs labor relations in Mexico.

2. Labor Law As Specific Domain

The labor justice system is responsible for effective enforcement of labor laws, together with other instances like the labor inspectorate and trade unions themselves. The malfunctioning of such a system, creates a strong incentive for employers to choose for not respecting the workers' rights, with the expectation that they will desist from claiming them facing the many obstacles that would have to overcome to implement them. Thus, one of the most common problems in our country is the lack of knowledge of laws concerning workers, that is, the information exists but is not divulged properly or adequately. Therefore, we have developed a system based on the search for answers that will allow the system to recover accurate information tailored to user's request, allowing you to always know the laws that apply to your specific case. In addition, we incorporate mobile technologies that can be put into the user's hand a tool that allows you to make your natural language queries and whose goal is to know his rights enshrined by law. Our main objective is modeling and development of a tool to assist in understanding and correct application of our Mexican labor law. With this novel proposal, we solve the needs that demand sectors such as: accounting firms, universities, offices, employees, etc.

3. State Of The Art

The general QA systems have focused all their energy to answer common sense questions. Namely, they often try to answer questions whose answer types are: date, person name, organization and so on. For instance,

When was Trec-10 held? Or Who was the President of USA?

Whose answers type are DATE and PERSON respectively. Recent research has focused on developing systems for question answering to open domain, that is, systems that takes as their source of information a collection of texts on a variety of topics, and solve questions whose answers can be obtained from the collection of departure. From question answering systems developed so far, we can identify three main phases [2]:

3.1. The Question Is Analyzed

This first phase will identify the type of response expected from the given question, that is expected to be a question of "When" a kind of response time, or a question "Where" will lead us to identify a place. Response rates are most commonly used personal name, name organization, number, date and place.

3.2. Recovery Of The Document

In the second stage performs a recovery process on the collection of documents using the question, which is to identify documents on the question that probably contain the kind of response expected. The result of this second stage is a reduced set of documents and preferably specific paragraphs.

3.3. Extraction Of The Answer

The last phase uses the set of documents obtained in the previous phase and the expected type of response identified in the first phase, to locate the desired response. Definition questions needs a more complex process in the third stage, since they must obtain additional information segments and at the same time are not repetitive. To achieve a good "definition" must often resort to various documents [2].

Currently the question answering on mobile devices for open domains, there are several commercial proposals how to we can see in [2] and [3].Seb Maslin founded the application shown on [2]. 199QUERY is the premier Australian and New Zealand Text service that answers any question sent to it by SMS using a combination of human experts and sophisticated algorithms to provide bespoke answers to any customer query. We note that applications 199query and AQA [2], which are not fully automatic because they require human assistance for its operation, as well as the use of such applications is limited, that is, you can count on their service only in countries where they operate. As opposed to our proposal, which is supported by intelligent agents for each of the techniques employed in our system. In next section, we present the architecture used for our mobile application.

4. Mobile Architecture

The development of mobile technology evolves vertiginously achieving change the behavior of societies radically. Just a few years ago, the penetration of mobile devices had only reached that users saw as entertainment devices and multimedia applications. However, these have evolved and grown from a simple consumers to be demanding customers, now not only used as a means of entertainment, but now are also socializing, buying, searching, creating content, etc., even now, these new users are referred to the term "prosumer" [4]. This term applies to users who act as channels of human communication, which means that at the same time be a consumer, are at the same time content producers. A "prosumer" has no lucrative purposes only participate in a digital medium of exchange of information; such is the case of peer-to-peer, networks interchangeable pairs. The word "prosumer" perfectly describes to million participants in the revolution of Web 2.0, because there are more and more people involved uploading information to the network and, at the same time are consumers of it, thereby creating a world of possibilities in all aspects. Furthermore, the mobile network is different, the bandwidth is narrow, and delays are greater than in the Personal Computer /Wired network, which supports 2 Megabits per second. The challenges for developing an application for the mobile environment are mainly these obvious differences between the Mobile network and the wired environment. Thus, the choice of infrastructure for application development is critical because response times are crucial to the acceptance thereof by the users. For this reason, we have designed an architecture that allows for processing hard on our server with intelligent and rapid methods that allow users to provide timely and accurate answers to your questions. Moreover, In Mexico charges 4 cents per KiloByte. Which gives us an adequate means to provide correct and accurate answers at low cost [5].

The proposed architecture for this application consists of two main modules: one focused on extracting the answers to the question given and other to send the response through a mobile device. The modules of the architecture system has the following functionalities:

4.1. Mobile Question Answering

Is the user interface for mobile device, which is responsible for communicating via GPRS (General Packet Radio Service), 3G o Wi–Fi with, web service definition question answering [5].

4.2. Web Service Definition Question Answering

Is the web service that satisfy all demands of mobile question answering, as well as requests the site and will have the task of looking for answer to questions of "definition". In figure 1, we can observe the architecture used in our proposal. With this architecture our server can attend all requests made from mobile devices and to each of these respond through the GPRS service. This architecture makes use of the gprs technology. Assumes a new switching network superimposed on the conventional GSM network. This technology is chosen for its high speed in data transfer. Furthermore, its availability on mobile devices is by default as well as its low cost.

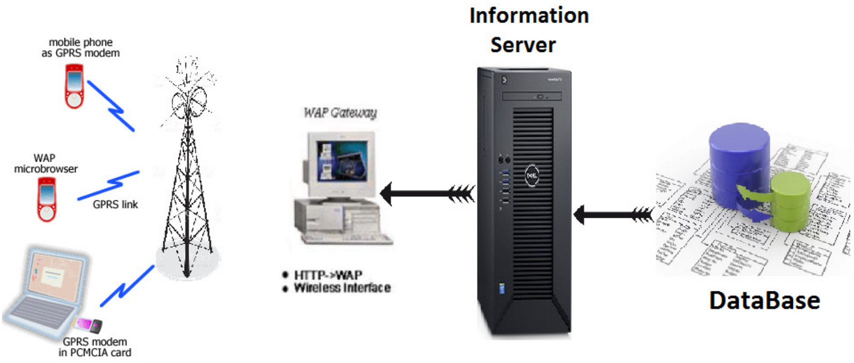


Figure1. Architecture for mobile application

Our method to question answering uses the Mexican Labor Law as collection of documents. We take advantage two qualities: 1) The Mexican Labor Law is organized into sections that are chapters, that is, a set of articles belonging to a section refers to a single topic and, 2) Each article has a unique small description that we call it issue. Thus, we can index our corpus based on these two characteristics, significantly improving recovery of relevant information. In general, our process to answer questions asked by users consists of three main modules that we will present in detail later.

5. Mobile Question Answering In Specific-Domains

The Mexican Labor Law are statutory provisions governing labor-management relations, that is, where are specified the rights and obligations to each one. In this law, there are 1010 articles, which talk about an issue specific. In our proposal we took advantage of the structure that labor law has, i.e., this has a brief general description for each of the articles that we used as passages. In a similar form, similarly, we exploit the fact that the articles are grouped into chapters. Finally, we use a specialized dictionary of labor terms.

Our algorithm is based on the integration of two databases described as follows:

First of these consists of five sections for each article, these sections are created with Lucene [6], which serve as indices in search for articles (answers), to know:

- The article's number
- The subject article formed by a brief description of issue addressed, as well as chapter to which it belongs
- The general description
- The combination, i.e., the combination of content and topic
- A thesaurus, that is, a list of specialized terms in the labor area. The second database is similar to the first, except that this will use lemmatization.

With respect to lemmatization, we developed our own lemmatisation algorithm truncating the last letters of the word to take only roots of the words. Finally, we define an efficient algorithm that allows us to obtain the best results in a fast and accurate way that satisfies the user's response.

Our algorithm is supervised by a main intelligent agent, which monitors the remaining agents embedded in each of the following steps:

1. If query in natural language references an article's number then, the agent specialized in this task is invoked to search in section of the article's number by the article required. To finally the agent deliver to the requesting user and the algorithm stop.
2. If step one does not apply then, our main agent analyze in more detail the query to select the agent who will transfer control. If the length is short then, the agent entrusted with section 2 (theme section) is invoked. Thus, this agent is responsible of applying the algorithm called "NEAR". "NEAR" tries to find that most of the terms are relatively close to each other and returns the answer, that is, words with diameter close. In another words, our algorithm return the most similar word based on root of the word.
3. Similar to step 2, if our main agent detects that not apply either step 1 or 2 then, the agent responsible for applying lemmatization is invoked. In this search as previous, we use lemmatisation of the words, but this function is simplest because it only seeks the equity terms comparing them.
4. If the lemmatization algorithm is not applicable then the principal agent activate the agent responsible for applying the database without lemmatization, whose goal is to find the correct answer to the question.
5. On the other hand, the main agent may choose to invoke the agent responsible for applying the specialized thesaurus. The use of a specialized thesaurus in a domain-specific QA system is essential. It helps determine the appropriate meaning of terms in the domain. For example, one of the common meanings of the term "accion" (in Spanish) "effect of making". However, in the legal domain, its meaning is "title that certifies and represents the value of each of those parts". In order to reduce ambiguity, some semantic information is added to this kind of term. In our processing of labor-law-related questions, the term "accion", will be tagged with the semantic category "title that certifies" which is a general concept in our thesaurus. The thesaurus that we have used is found in [7].
6. Finally, as final strategy we use a function called "FUZZY", in this function we use the Levenshtein algorithm. The Levenshtein algorithm is described as distance of editing, that

is, the minimum number of operations required to transform one string characters in another can be an insertion, deletion or substitution of a character, for this, we use a dictionary of words. In this way we are sure to transform a natural language word for another that whether this contained in our database.

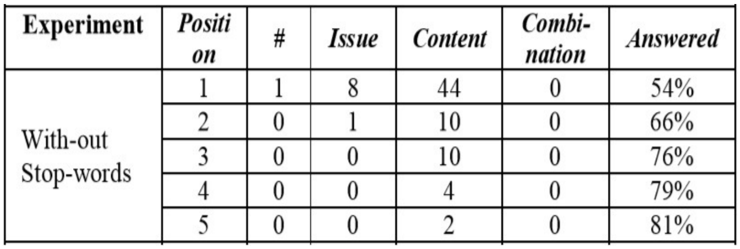
Thus, our algorithm finds the best response to the user, i.e., our algorithm deliver: the article(s) request(s), articles related to the question asked by the user, the passage nearest to article(s) requested, but in all cases, it delivers a concrete and correct answer.

6. Findings

In this section we present the evaluation that we have made to our system in order to have a clear perspective of system efficiency. For this, we have developed a series of tables, which present the results obtained with our system. To test our system we consider a corpus of 100 questions from students and professors of the faculty of the Accounting School at Universidad Autónoma de Puebla.

As shown in Table 1 the results obtained by agent that apply the methodology called without Stop-words are acceptable, considering that in this case, our algorithm removes only empty words such as: preposition, conjunctions, question words, articles, etc. Likewise, all words are eliminated whose repetition frequency is high, some of these are work, workers, article, boss, etc., as used in [8]. Thus, our proposal gives an added value to our application, since it allows users to make their requests without worrying about their writing.

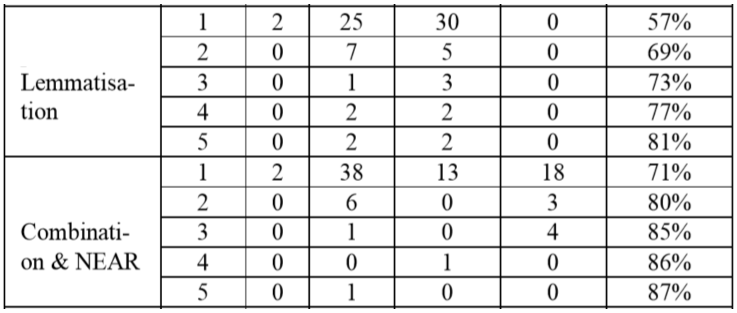
Table 1. Results With-out Stop-Words



Experiment	*Positi on*	#	*Issue*	*Content*	*Combi-nation*	*Answered*
With-out Stop-words	1	1	8	44	0	54%
	2	0	1	10	0	66%
	3	0	0	10	0	76%
	4	0	0	4	0	79%
	5	0	0	2	0	81%

Next, in Table 2 the results obtained are shown two strategies called lemmatisation and near. In this case, the results are better than previous in 15%, this means that our proposal improves when our agents combine their results. This is one of the advantages provided by the use of intelligent agents. This is because an agent decide whether it is desirable or not to use a more sophisticated strategy. This decision is based on the certainty of results, ie, if they are very low (55% to 65%) then used some of these strategies and even the combination of these. It is important to note that the use of intelligent agents allows correct decision making.

Table 2. Lemmatisation & Near



Lemmatisa-tion	1	2	25	30	0	57%
	2	0	7	5	0	69%
	3	0	1	3	0	73%
	4	0	2	2	0	77%
	5	0	2	2	0	81%
Combinati-on & NEAR	1	2	38	13	18	71%
	2	0	6	0	3	80%
	3	0	1	0	4	85%
	4	0	0	1	0	86%
	5	0	1	0	0	87%

Finally, in table 3 our results are presented, which as you can see are very good. Thus, at the bottom of the table 3, it presents the best results obtained by our system. Of 100 questions raised in natural language to our system, it answer 81% of these, that is of 10 questions raised to our system, this answer 8 correctly in the first instance. And if we allow the system give 3 answers to each question, this increases their effectiveness to 90%, improving the results reported by other similar systems. This result is excellent for this type of systems, considering that the terminology used is very technical. This means that we can consider that our system is as effective as a specialist advisor. In addition, if we look more instances, our system achieves a certainty of 92%, making it even better than as any human does.

Table 3. Satisfactory results obtained by our proposal

Levenshtein	1	2	38	13	22	75%
	2	0	5	0	2	82%
	3	0	1	0	2	85%
	4	0	0	1	0	86%
	5	0	1	0	1	87%
abbreviation dictionary	1	2	38	14	22	76%
	2	0	4	0	4	84%
	3	0	0	1	0	85%
	4	0	0	0	1	86%
	5	0	1	0	1	88%
chapter paragraph	1	2	41	14	24	81%
	2	0	2	0	3	86%
	3	0	2	0	2	90%
	4	0	0	1	0	91%
	5	0	1	0	0	92%

7. MOBILE SYSTEM FOR QA

The rapid evolution of technology, especially focused on mobile devices and wireless network, has allowed more people to acquire these services. Now is not necessary to have a computer and an internet provider at home, mobile data coverage of different companies has begun to replace this service. If also added that mobile phones have the GPRS service included in the WAP protocol, this makes it more attractive than an application run on these.

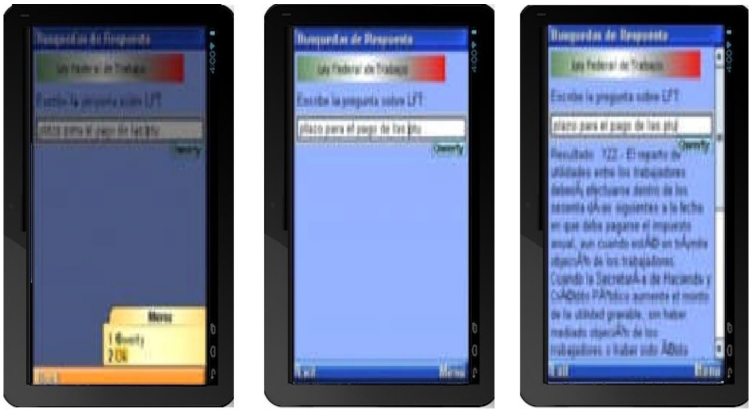


Figure 2. Mobile application for Question Answering

In figure 2, the interface of mobile application developed to answer questions related to Mexican labor law in natural language is deployed. This gives an added value to our proposal. This application was developed in Spanish language, because it is used in Mexico and the Mexican labor law is in Spanish, in addition, users are Spanish speaking. The architecture used for the development of this application is similar to that presented in [5], [8] and [9]

8. CONCLUSIONS

We have developed a mobile system for question answering (mQA) in the context of the Mexican labor law, achieving a number of successes 92%, and a response time between 12 and 20 seconds. However, the most remarkable aspect is effectiveness achieved by our system. We have developed the basis for the implementation of a new methodology to answer questions in natural language for specific domains.

Consider that our system for search for answers in domain-specific is a very useful tool in everyday life, and that any user may take advantage of their information to certain difficulties that arise, as well as the great interactivity that this provides to Question Answering systems. Also, consider the Lucene system as a useful tool in extracting passages, versatility and flexibility provided by the tool makes the task of searching for question answering is simpler and more accurately.

As future work, we can incorporate other dictionaries of terms to be used as synonyms for words infrequently. With this, you can get more certainty and achieve 100% effectiveness. Furthermore, we can implement a method of validation of answers to questions about the Mexican Labor Law. Finally, we can use "InQuery" for the extraction of passages and compare it with ours.

ACKNOWLEDGEMENTS

The authors wish to thank The Universidad Autónoma de Puebla for their financial support. In addition, we appreciate support of the academic body Combinatorial Algorithms and Learning.

REFERENCES

[1] PewResearchCenter (2013). Internet, Science & Tech. Informe de Pew Internet. Teens and technology 2013. http://www.pewinternet.org/2013/03/13.

[2] Burger John, Cardie Claire, Chaudhri Vinay, Gaizauskas Robert, Harabagiu Sanda, Israel David, Jacquemin Christian, Lin Chin-Yew, Maiorano Steve, Miller George, Moldovan Dan, Ogden Bill, Prager John, Rilo+ Ellen, Singhal Amit, Shrihari Rohini, Strzalkowski Tomek, Voorhees Ellen, Weishedel Ralph. Issues, Tasks, and Program Structures to Roadmap Research in Question Answering (QA). Technical Report, National Institute of Standards and Technology.

[3] Bakker, Dik, André Muller, Viveka Velupillai, Soren Wichmann, Cecil, H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. (2009). Adding typology to lexicostatistics: a combined approach to language classification. Linguistic Typology 13: 167-179.

[4] Junyeon Hwang, Myeong in Choi Tacklim Lee Seonki Jeon Seunghwan Kim Sounghoan Park Sehyun Park. Energy Prosumer Business Model Using Blockchain System to Ensure Transparency and Safety. Volume 141, December 2017, Pages 194-198. Energy Procedia, Elsevier.

[5] Fernando Zacarias, Rosalba Cuapa, Antonio Sanchez and Iris Cerecedo (2011). Puebla in the palm of your hands. MoMM 2011: 260-263, ACM New York, USA. ISBN: 978-1-4503-0785-7.

[6] Doug Cutting. (1999). http://lucene.apache.org/.

[7] Nava Tovar A. (2015) Diccionario Jurídico, La institucionalización de la razón. Universidad Autónoma Metropolitana. http://www.diccionariojuridico.mx/. Anthropos Editorial.

[8] Balderas Espinosa M.A. (2008) Master thesis. Faculty of Computer Science at Universidad Autónoma de Puebla.

[9] Fernando Zacarias, Rosalba Cuapa, Guillermo De Ita, J.C. Acosta and Daniel Torres (2014). Planning Solutions in the Real World. International Journal of Artificial Intelligence & Applications. Vol. 5 N0. 3. IJAIA –AIRCC Publishing Corporation

[10] Zhuo, Lyne, Colin, Gonzalo, Jian (2004). Domain-specific QA for the Constructor sector.

AUTHORS

He is researcher and professor of computer science at the Universidad Autónoma de Puebla. He is a researcher in practical and theoretical computer science and mobile technologies. He has conducted R&D projects in this area since 2000. Results from these projects have been reported in more than 60 national and international publications. Professor Zacarias serves at the editorial board of the Journals: IEEE Latin America Transactions, Engineering Letters, International Transactions on Computer Science and Engineering, Common Ground Publishing - Technology, Learning and Social Sciences.

Guillermo De Ita Luna He did his BS in Computer Science in the Faculty of Computer Sciences in the Autonomus University of Puebla (BUAP), México. The master and Ph.D. Programing Electrical Engineering in the Cinvestav - I.P.N., México. He has worked by 10 years as developer and consulter for Database Systems and Geographic Information Systems for different enterprises in México.