COMUNICA - A Question Answering System for Brazilian Portuguese

Rodrigo Wilkens[♣], Aline Villavicencio[♣], Daniel Muller[◇], Leandro Wives[♣], Fabio da Silva[♠], Stanley Loh[♡]

[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil),
[◇]Conexum (Brazil), [♠]DFL (Brazil), [♡]IntextMining (Brazil)

`{rwilkens,avillavicencio,wives}@inf.ufrgs.br, conexum@conexum.com.br, fabio@dfl.psi.br, sloh@terra.com.br`

Abstract

COMUNICA is a voice QA system for Brazilian Portuguese with search capabilities for consulting both structured and unstructured datasets. One of the goals of this work is to help address digital inclusion by providing an alternative way to accessing written information, which users can employ regardless of available computational resources or computational literacy.

1 Introduction

A crucial social problem in many countries is functional illiteracy, and in Latin America, according to UNESCO, the process of literacy is only effectively achieved for people who completed at least four years of schooling. Among those who have not completed this cycle of education, there has been high rates of return to illiteracy. According to this definition, in 2002 Brazil had a total of 32.1 million functionally illiterate citizens, representing 26% of the population aged 15 or older[1]. This may have a significant effect on digital inclusion, preventing a considerable part of the population from accessing massive amounts of information such as that available on the Web, or benefitting from advances in technology. Although these figures do not include digital iliteracy, or lack of computational resources, they can give an idea of the magnitude of the problem.

In this context, voice question answering systems (QA) have the potential to make written information more easily accessible to wider audiences as they allow users to ask questions in their own native language and especially if this includes spoken language, sometimes without the need even for a computer (e.g. using the phone). This paper describes COMUNICA, a voice QA system for Brazilian Portuguese with search capabilities for consulting both structured and unstructured datasets. The domain chosen to evaluate the system is that of municipal information from the FAMURS database.[2] One of the goals of this work is to help address digital inclusion by providing a way to overcome (a) difficulties in accessing written information (for visually challenged users), (b) lack of computational resources (for users in remote or computerless areas) and (c) computational illiteracy.

2 QA systems

In recent years, QA has received considerable attention, as can be seen by the initiatives devoted to the task, such as the TREC[3] and CLEF[4]. The task of a QA system is to automatically answer a question in natural language, searching for information in a given data source (e.g. a database, or corpora from a given domain). This is a challenging task as question types can range from lists to facts and definitions, while answers may come from small data sets such as document collections, to the World Wide Web. Moreover, the difficulty of the task is also influenced by whether the questions are restricted to a particular domain (e.g. sports, genes) or not, which additional sources of in-

[1] IBGE: http://www.ibge.gov.br/ibgeteen/pesquisas/educacao.html

[2] `http://www.famurs.com.br`

[3] `http://trec.nist.gov`

[4] `http://www.clef-campaign.org`

Coling 2010: Demonstration Volume, pages 21–24,
Beijing, August 2010

formation are available for a given language (e.g. ontologies of domain-specific knowledge, general ontologies), their coverage, and which tools can be used to help the task (e.g. named entity recognisers, parsers, word sense disambiguation tools). Furthermore, there is no concensus as to the amount of resources and tools that are needed in order to build a working QA system with reasonable performance.

For a resource rich language like English, there is a consistent body of work exemplified by systems such as JAVELIN (Nyberg et al., 2002) and QuALiM (Kaisser, 2005). For other languages, like Portuguese, and particularly the Brazilian variety, QA systems are not as numerous. Over the years, there was an increase in the number of participating systems and data sources in the CLEF evaluation. For instance, in 2004 there were 2 participating systems, and in 2006 it had 4 systems and the best performance was obtained by Priberam (Amaral et al., 2005) with 67% accuracy (Magnini et al., 2006). Figure 1 summarizes the performance of the QA systems for Portuguese for QA@CLEF over the years.

3 COMUNICA Architecture

The Comunica system is composed of five modules: a manager module and four processing modules, as shown in figure 2. The manager is responsible for the integration and communication with the speech recognition, text processing, database access, and speech synthesis modules.

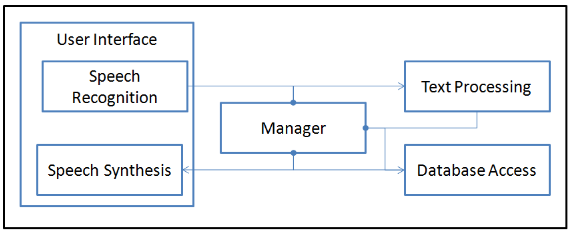

Figure 2: Architecture of the system.

3.1 Speech Recognition

For continuous speech recognition of the users' requests we use an automated phone service. This module uses two research fronts signal analysis (Fourier transform and Wavelets). The coefficients obtained are sequenced on three fronts for continuous speech recognition: HMMs (Becerikli and Oysal, 2007) TDDNN and NESTOR (Nasuto et al., 2009). To train the models, a corpus of FAMURS callcentre telephone interactions has been recorded. The recognition focuses on the vocabulary employed in the domain, in this case municipal information related to taxes from FAMURS. In order to do that, it uses 2 ontologies to validate the candidate words in the input: (a) a general purpose and (b) a domain ontology. The recognised transcribed input is passed to the manager for further processing.

3.2 Text Processing

The manager sends the transcribed input to be processed by the natural language processing module. The natural language queries are processed using shallow and deep tools and accessing both a general and a domain specific ontologies (illustrated in Figure 3). This module needs to determine which type of query the user performed and what is the likely type of answer, based on mostly lexical and syntactic information. This process is divided into 3 mains steps: parsing, concept identification and pattern selection. In the first step, the input is parsed using the PALAVRAS parser (Bick, 2002), and the output provides information about the particular pronoun (wh-word), subject and other verbal complements in the sentence. For concept identification, the system uses the domain ontology, which contains the relevant concepts to be used in next steps. The ontologies also provide additional information about nouns (such as hyperonymy and synonymy) for determining which instances of the concepts were present in the input. For example, "Gramado" is an instance of

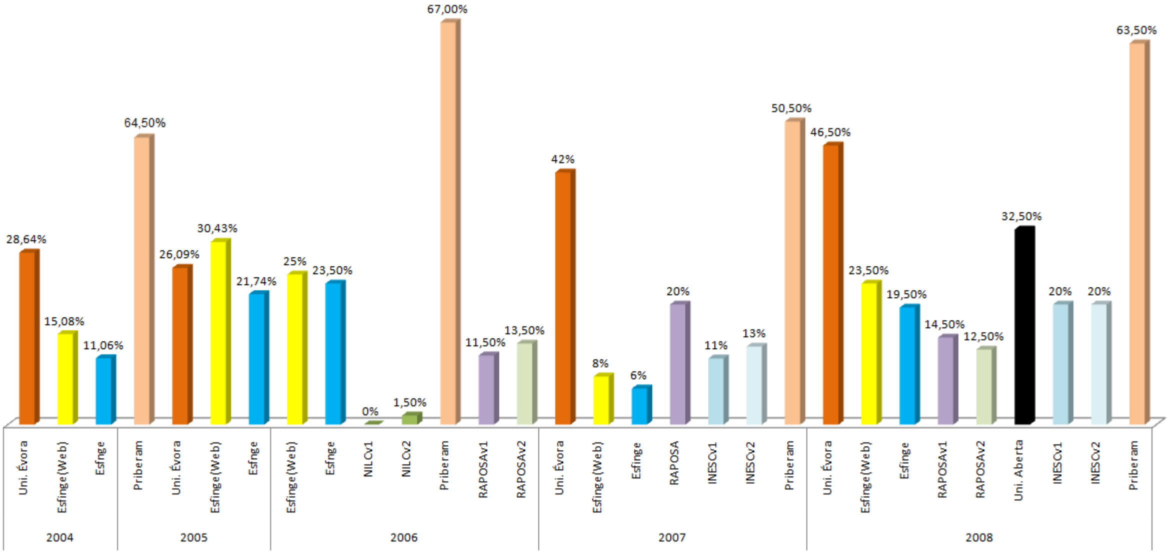

Figure 1: Performance of QA systems for Portuguese QA-CLEF.

the concept "city". Both absolute and relative dates and periods (e.g. last quarter, first week) need to be treated.

Finally, based on this information this module selects from a set of pre-defined question patterns linking concepts of the domain ontology with SQL commands, the one which contains the largest number of concepts in common with the input, and sends it to the manager in an XML format. If there is no complete frame, this module identifies which concepts are missing and returns this in the XML output.

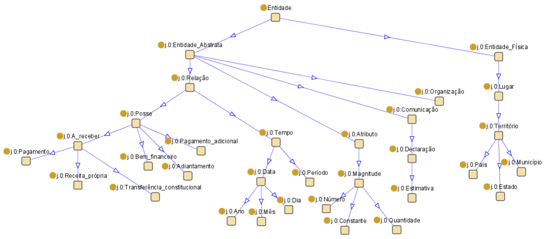

Figure 3: The domain ontology

3.3 Database Access

The search module is divided in two submodules: one for searching information in a structured database and the other for searching in an unstructured knowledge base. It receives as entry an XML file, containing the original input in natural language and the concepts identified in the question. The structured search module receives the input tagged with concepts of the ontology and an identified search pattern, and selects a structured SQL query. These queries are predefined according to the search patterns and the structure of the database. For example, in the case of the FAMURS domain, there are concepts related to time period, cities and taxes. When these 3 concepts are found in the input, a special pattern is selected which defines the kind of information that must be retrieved from the database. An SQL command is then executed in the structured database. All possible patterns are mapped to a specific SQL command. These commands have slots that are filled with instances of the concepts identified in the sentence. For example, names of cities are instances of the concept "city". The retrieved values are used for producing the answer in natural language, using some predefined answer patterns.

Otherwise, the system uses the ADS Digital Company Virtual Assistant (VA) (Duizith et al., 2004) to search the unstructured data (e.g. Frequently Asked Questions), using the lexical information to locate the answer associated to the most similar question. This answer is written in natural language and will be returned to the main module of the system. If no similar question is found according to a predefined degree of similarity, the VA returns a standard answer.

3.4 Speech Synthesis

The text output to the user is synthesized, resulting in an audio file that is transmitted through the server.

3.5 Manager

The manager is responsible for the integration and communication of the modules. It processes requests, interpreting the actions to be taken and dispatching the requests to specific modules. To start the interaction the manager activates the speech recogniser, and if no problem is detected with the input, it is passed to to the text processing module. In the case of missing information, the manager informs the user that more information is needed. Othwerise, the query is passed to the database module. The database module then returns the result of the query to the manager, which sends this information to the interface component.

All the components are SOA compliant and designed as Web services. This allows us to use a common and simple way of communication among components, allowing a certain degree of independence. Then components can be implemented using different technologies and may be distributed among different servers, if needed.

4 System Demonstration

This is an ongoing project, and a working version of the system will be demonstrated through some text example interactions from the FAMURS domain as the speech recognizer and synthesizer are currently under development. However, users will be able to interact with the other modules, and experience the benefits of natural language interaction for accessing database information.

Acknowledgments

This research was partly supported by CNPq (Projects 479824/2009-6 and 309569/2009-5), FINEP and SEBRAE (COMUNICA project FI- NEP/SEBRAE 1194/07).

References

Amaral, Carlos, Helena Figueira, André F. T. Martins, Afonso Mendes, Pedro Mendes, and Cláudia Pinto. 2005. Priberam's question answering system for portuguese. In Peters, Carol, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *CLEF*, pages 410–419. Springer.

Becerikli, Yasar and Yusuf Oysal. 2007. Modeling and prediction with a class of time delay dynamic neural networks. *Applied Soft Computing*, 7:1164–1169.

Bick, Eckhard. 2002. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. Ph.D. thesis, Aarhus University.

Duizith, José Luiz, Lizandro Kirst da Silva, Daniel Brahm, Gustavo Tagliassuchi, and Stanley Loh. 2004. A virtual assistant for websites. *Revista Eletronica de Sistemas de Informação*, 3.

Kaisser, Michael. 2005. Qualim at trec 2005: Web-question answering with framenet. In *Proceedings of the 2005 Edition of the Text REtrieval Conference*, TREC 2005.

Magnini, Bernardo, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Valentin Jijkoun, Petya Osenova, Anselmo Peñas, Paulo Rocha, Bogdan Sacaleanu, and Richard F. E. Sutcliffe. 2006. Overview of the clef 2006 multilingual question answering track. In Peters, Carol, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *CLEF*, volume 4730 of *Lecture Notes in Computer Science*, pages 223–256. Springer.

Nasuto, S.J., J.M. Bishop, and K. DeMeyerc. 2009. Communicating neurons: A connectionist spiking neuron implementation of stochastic diffusion search. *Neurocomputing*, (72):704–712.

Nyberg, Eric, Teruko Mitamura, Jaime G. Carbonell, James P. Callan, Kevyn Collins-Thompson, Krzysztof Czuba, Michael Duggan, Laurie Hiyakumoto, N. Hu, Yifen Huang, Jeongwoo Ko, Lucian Vlad Lita, S. Murtagh, Vasco Pedro, and David Svoboda. 2002. The javelin question-answering system at trec 2002. In *TREC*.