Robust Question Answering

Gracinda Carvalho[1,2,3], David Martins de Matos[2,4], and Vitor Rocio[1,3]

[1] Universidade Aberta, Rua da Escola Politécnica, 147 1269-001 Lisboa, Portugal
gracindac@uab.pt,vjr@uab.pt
[2] L2F/INESC-ID Lisboa, Rua Alves Redol 9 1000-029 Lisboa, Portugal
david.matos@inesc-id.pt
[3] CITI - FCT/UNL
[4] Instituto Superior Técnico/UTL

Abstract. A Question Answering (QA) system should provide a short and precise answer to a question in natural language, by searching a large knowledge base consisting of natural language text. The sources of the knowledge base are widely available, for written natural language text is a preferential form of human communication. The information ranges from the more traditional edited texts, for example encyclopaedias or newspaper articles, to text obtained by modern automatic processes, as automatic speech recognizers.

The work developed in the present thesis focuses on the Portuguese language and open domain question answering, meaning that neither the questions nor the texts are restricted to a specific area, and it aims to address both types of written text. Since information retrieval is essential for a QA system, a careful analysis of the current state-of-the-art in information retrieval and question answering components was conducted. A complete, efficient and robust question answering system is developed in this thesis, consisting of new modules for information retrieval and question answering, that is competitive with current QA systems. The system was evaluated at the Portuguese monolingual task of QA@CLEF 2008 and achieved the 3rd place in 6 Portuguese participants and 5th place among the 21 participants of 11 languages.

The system was also tested in Question Answering over Speech Transcripts (QAST), but outside the official evaluation QAST of QA@CLEF, since Portuguese was not among the available languages for this task. For that reason, an entire test environment consisting of a corpus of transcribed broadcast news and a matching question set was built in the scope of this work, so that experiments could be made. The system proved to be robust in the presence of automatically transcribed data, with results in line with the best reported at QAST.

1 Introduction

1.1 Description of the Problem and Motivation

The purpose of a Question Answering (QA) system is to provide an answer, in a short and precise way, to a question in Natural Language. Answers are produced by searching a knowledge base that usually consists of Natural Language text. The usefulness of this type of system is to find the exact information in large volumes of text data. With the wider availability of this type of resources, whether it is in the form of newspaper collections, or texts obtained through ASR (Automatic Speech Recognition), or encyclopaedic resources, the blogosphere, social networks or even the World Wide Web, there is an increasing interest in this type of system.

We dedicate our attention to open-domain question answering, as described in [1]. In recent years there has been very active research in this field, that led to the creation of tracks dedicated to QA in several international evaluation initiatives as is the case of TREC of the Text REtrieval Conference (TREC), for the English language, NTCIR Workshop whose main focus is on Asian languages, and the Cross-Language Evaluation Forum (CLEF), an initiative co-sponsored by the European Commission running a QA track, QA@CLEF, since 2003, including the Portuguese language since 2004.

Questions addressed by the current state-of-the-art include questions covering factual information of several types (e.g. Qual é a área da Groenlândia?[What is the area of Greenland?]), definitions (e.g. O que é o jagertee? [What is jagertee?]), list questions (e.g. Por que estados corre o Havel? [For which states does the Havel run?])and cluster questions , i.e. groups of questions linked by anaphoric references (e.g. Quem foi o criador de Tintin? [Who was the creator of Tintin?] and Quando é que ele foi criado? [When was he created?]), and they may include temporal restrictions (e.g. Quantos habitantes tinha Berlim em 1850? [How many inhabitants did Berlin have in 1850?]). The examples given all belong to the Portuguese monolingual task of QA@CLEF 2008.

1.2 Thesis Objectives and Approach

The techniques employed by current state-of-the-art Information Retrieval (IR) and Question Answering (QA) systems are investigated in the thesis and subject to a careful inspection of practical aspects, as well as of their theoretical motivations. Since we believe that a system can never be too simple, as long as it complies with the specifications and produces good results, we make a careful analysis of the cost/benefit of the techniques liable to be employed, valuing simpler solutions.

The main goal of this work is to study and develop innovative components of IR and QA, to build a complete, efficient and robust QA system, that can compete with the current state-of-the-art QA systems. The name of the system developed is IdSay, a short name for ”I would say” or ”I dare say”.

The key concepts taken into account in the development of the present work are:

- efficient implementation;
- robust implementation;
- validation of the results through evaluation with peer systems;
- explore Wikipedia as a resource for QA;

– use developed QA system on speech transcripts.

If the system is not efficient, it will not be useful for a real question answering application, since users expect fast answers. We are only interested in fast components of IR and QA, that are less likely to give enough time for the user to consider the decision of waiting for the answer or giving up.

A robust system, that performs well not only under the most favourable conditions but also under unusual circumstances, has greater chances of being valuable in more situations, so we will direct the development towards the robustness of the components. In our case, the most favourable circumstances refer to well formed text, as opposed to text that contains incorrections such as that obtained through automatic methods.

It is extremely important to validate the system in an international forum, not only to be able to compare it with peer systems, and to share the evaluation effort with others, but mainly to have the results certified by an international organization, that prevents misleading analysis and conclusions on the performance of the system. Therefore, whenever possible, we use this option.

An important option taken is to use the Portuguese Language, a widely spoken language, which makes it both eligible in terms of the extensive text data present for instance in the web and also the usefulness for a large number of potential users. Despite this fact, language resources are still less abundant, especially if we make a comparison to English, the current "standard language" for research, but also to some other languages. However facts have proven that it is possible to achieve state-of-the-art results using Portuguese, compensating the possible lack of specific resources.

Wikipedia combines two characteristics that we consider interesting, one is the fact that it is freely available for public use, and the other is the fact that it results from a collaborative effort from millions of people around the globe, bringing it the benefits of diversity and volume. Both these features contribute for the creation of quality working material. We intend to make use of Wikipedia for improving the system efficiency.

Finally we want to test the robustness of the system using it on data obtained from Automatic Speech Recognition (ASR) applications. Because this data is less well formed than written text, due to the word error rates (WER) of the recognisers, a good performance in this scenario validates the robustness of the system. It is an important application of QA since more data from ASR is becoming available, along with the corresponding need for search mechanisms to cope with it.

1.3 Document Organization

Sect. 2 is dedicated to the steps followed in the development of IdSay system, including its baseline version that was evaluated at QA@CLEF 2008, and improvements that were made after the analysis of its results, as well as those of other systems participating in the Portuguese monolingual task. In Sect. 3 we describe the experiments with speech transcripts used to validate the robustness of the system. To finalize, in Sect. 4 we present our conclusions and future work.

2 IdSay

In the present section we describe the architecture and main features of IdSay system.

The core version of IdSay uses little linguistic knowledge, and is as close as possible to simple keyword search. The only external information that we use besides the text collections is lexical information for Portuguese [2].

Our option for using little linguistic knowledge is to maintain the system less dependent of a specific language and its rules, as well as to be able to use the system with noisy text, i.e. text with is not fully compliant with the rules of the language. Also the language processing tools sometimes are very time consuming, so the whole process is slowed down.

The architecture for IdSay is presented in Fig. 1.

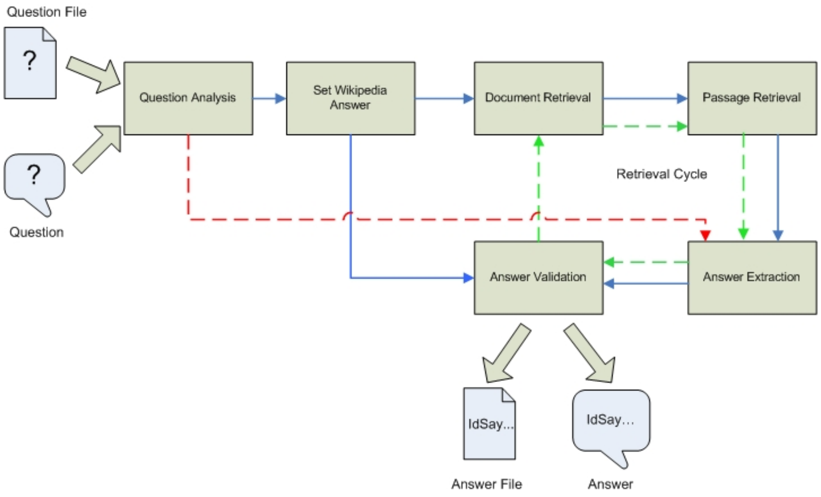

Fig. 1. IdSay system architecture

IdSay accepts either a question written by the user (manual interface), or a set of questions in an XML file (automatic interface). In the manual interface, the system is not prepared to treat co-reference between questions.

Each question is treated in the question analysis module to determine the question type and other information to be used in answer extraction module (red dashed line in Fig. 1). The question analysis also determines a search string with the information of which words and entities to use in the document retrieval module to produce a list of documents that match both. In special cases, for instance definition questions about an entity that has a Wikipedia page, the Set Wikipedia ANswer (SWAN) module produces an answer based on that page, that is directly included in the answers to be treated in the Answer Validation module. In the general case the process proceeds along the blue line in Fig. 1, to the document retrieval module, in which the documents of the collection are searched based on the words and entities information derived from the question, producing the list of documents that contain all of them. This list of documents is then processed by the passage retrieval module, responsible for the search of passages from the documents that contain the search string, and with length up to a given limit. The passages are then sent to the answer extraction module, where short segments of text (candidate answers) are produced, that are then passed on to the answer validation module . This module validates answers and returns the most relevant ones. If in one of the steps no data is produced, the search string is revised and the loop starts again, in a process we identify as retrieval cycle (green dashed line in Fig. 1).

This is a classic QA system architecture, but contrary to most QA systems, we do not store passages in the IR module, but documents. Passages are extracted in real time, depending on the question information. This option allows more flexibility, and it is especially suited for the case of text obtained from speech transcripts, in which sentences are not clearly defined, as in manually written and edited text. The retrieval cycle introduces a question based mechanism for leaving out words when results produced so far are not yet satisfactory.

A baseline of the system was submitted to evaluation at the 2008 edition of QA@CLEF that obtained an accuracy of first answers of 32.5% acorresponding to the 3rd position among the 6 participating systems in the same task, in which it was the only system participating for the same time [3]. This result allowed IdSay to be placed in the 5th position among the 21 participating sytems in the 11 languages offered [4]. The good results achieved for the Portuguese language is reflected by the fact that the first position was obtained by the Portuguese company Priberam with an an accuracy of first answers of 63.5% and the 3rd place was obtained by Universidade de Évora with an accuracy at first of 46.5%.

We built a web application[5] based on the version of IdSay whose results were submitted to QA@CLEF 2008. The purpose of this web application is manifold: this way we are able to reproduce on-line the results obtained at the CLEF campaign, but we can also use the system for any other question. The added debug options enable a deep analysis of the results obtained.

The improvements introduced after the analysis of the results obtained in the evaluation campaign (by IdSay, as well as all other participating systems for Portuguese) allowed the results to be improved to an accuracy of 50.5%. These improvements were achieved especially by the introduction of equivalences between words, using the TeP base [5]and introduction of equivalences at entity level, by means of a resource WES base (Wikipedia Entity Synonymns) that we built automatically based on information from Wikipedia, following the TeP base format, to allow integration of both resources. Altough the resource was produced for Portuguese, it is language independant, depending only on the language of the Wikipedia being used. A framework of analysis for systems

[5] Available at http://www.idsay.net.

performing the same task, the results quadrants was proposed and used for the case ofthe systems participating at QA@CLEF 2008 for Portuguese [6].

IdSay system relies on an efficient search and indexing component, IdSearch, also fully developed in the scope of the thesis, and it takes in the improved version only 4 hours to index the QA@CLEF 2008 collection (9.5 GB of text, including 2 years of newspaper articles form a Portuguese and a Brazilian newspaper, together with a frozen version of the Wikipedia), and about 2 minutes to answer the to the 200 questions in the question set, which represents a mean answering time of 0.6 seconds per question. None of the participants of QA@CLEF published results concerning their efficiency, with results being published only for the case of a system that did not participate in the evaluation, that reports a mean answering time of 22 seconds per question [7].

3 IdSay on Speech Transcripts

The robustness of the system was tested in a case study related to questions on speech transcripts. The formal evaluation at QA@CLEF, called QAST (Question Answer on Speech Transcripts) was not an option since it included only the English, French and Spanish Languages. For that purpose we built a corpus for the Portuguese Language and a corresponding question set.

The data collection consists of video recordings of the evening editions of the Broadcasting News from the two channels of the Portuguese public television network, Rádio Televisão Portuguesa, RTP along with a set of 100 questions. The editions of the Broadcast News shows are from the 1st of June to the 11th September of 2008, a period that included the Olympic Games of Beijing 2008, and it amounts to a total of 103 days and over 206 shows. The data contains over 180 hours of audio with the corresponding automatic transcripts [8], enriched with punctuation marks [9]. Approximately 60 excerpts were transcribed manually, and a set of 100 questions was made based on these transcripts.

The system proved its robustness, for even in the presence of text with words incorrectly transcribed, or misplaced punctuation marks the system was able to provide the correct answer in the first place for 30% of the questions, 42% of questions considering the first three answers returned, and an accuracy of over 60% in the location of the passage that contained the correct answer, in the best setup for the experiments. This last value is interesting for an application that retrives the correct video excerpts in a collection of may hours of video, based on a question stated in Natural Language. The results are in line with the best obtained at QAST [10].

4 Conclusions and Future Work

4.1 Contributions

The major contribution of the present work is the design and development of an innovative QA system, IdSay, with focus on efficiency and robustness. We highlight the most relevant contributions of the present thesis:

- A statistically validated study was conducted regarding pre-processing options for an IR system working in the QA context for Portuguese [11]. We found out that converting text to lowercase and removing punctuation marks increase retrieval efficiency but there is no statistical evidence of improvements derived from the use of stop lists, lemmatization or stemming, for the experiments conducted. These results showed that despite common assumptions of relevance of these techniques in the literature, their use does not automatically lead to an improvement in IR accuracy.
- A method of Results Quadrants is introduced, that summarizes the information related to a system when compared to other systems performing the same task, in this case answering questions. The Results Quadrant for a system allows the identification of such characteristics as its degree of innovation or coverage of easy questions, in perspective with peer systems.
- An efficient search mechanism for large document collections to be used for QA (IdSearch). The data structure for storing documents uses one number per word, instead of strings. This data structure, in the text collection used, requires an average of 4,28 bytes/word while the string version would require 10,43 bytes/word. With this data structure, instead of string manipulation one integer comparison is done to compare words. This leads to improvements in both space and time. The component is based on the Boolean Retrieval Model as the result of the study conducted, and the search is done considering separately words and entities(groups of contiguous words) identified from the question.
- A strategy to remove the most frequent keywords from the query, only if no satisfactory results have been produced. This can be seen as a dynamic application of stop lists, but instead of blindly removing stop words at indexing time, the words are selectively removed, if needed, based on the question being processed.
- A Question Answering over Speech Transcripts corpus (described in 3) for Portuguese, which was used to test successfully IdSay's robustness.

4.2 Further Areas of Application

We believe in the utility of both the aim of QA systems and our approaches in the context of growing volume of unstructured text information namely in the web. This text information comes in different languages, so to address the problem of aiding the users in accessing that information it makes sense to develop multi-language systems.

Another future research direction to follow emerges from the case study and has to do with search in audio/video streams using textual information obtained automatically.

Finally, and given the author's professional functions and interest in teaching and on-line learning, to explore the educational potential of QA constitutes another area of application to be explored, through the creation of automated, innovative methodologies and tools designed to promote the students' learning process.

References

1. Prager, J.M.: Open-Domain Question-Answering. Foundations and Trends in Information Retrieval **1** (2006) 91–231 DOI: http://dx.doi.org/10.1561/1500000001.
2. Alves, M.A.: Engenharia do Léxico Computacional: princípios, tecnologia e o caso das palavras compostas. Master's thesis, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Lisboa, Portugal (2002)
3. Carvalho, G., de Matos, D.M., Rocio, V.: IdSay: Question Answering for Portuguese. In: Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, LNCS Series Volume 5706, Springer-Verlag, Berlin, Heidelberg (2009) 345–352 DOI: http://dx.doi.org/10.1007/978-3-642-04447-2_40.
4. Forner, P., Peñas, A., Agirre, E., Alegria, I., Forăscu, C., Moreau, N., Osenova, P., Prokopidis, P., Rocha, P., Sacaleanu, B., Sutcliffe, R., Sang, E.T.K.: Overview of the CLEF 2008 Multilingual Question Answering Track. In: Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Springer-Verlag, Berlin, Heidelberg (2009) 262–295 DOI: http://dx.doi.org/10.1007/978-3-642-04447-2_34.
5. Maziero, E.G., Pardo, T.A., Felippo, A.D., Dias-da-Silva, B.C.: A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Electrônico para o Português do Brasil. In: VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL). (2008) 390–392
6. Carvalho, G., de Matos, D.M., Rocio, V.: Improving IdSay: a characterization of strengths and weaknesses in Question Answering systems for Portuguese. In: Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010, LNCS Series Volume 6001, Springer-Verlag, Berlin, Heidelberg (2010) 1–10 DOI: http://dx.doi.org/10.1007/978-3-642-12320-7_1.
7. Branco, A., Rodrigues, L., Silva, J., Silveira, S.: XisQuê: An Online QA Service for Portuguese. In: Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language, PROPOR 2008, Curia, Portugal, September 8-10, 2008, LNCS Series Volume 5190, Springer-Verlag, Berlin, Heidelberg (2008) 232–235 DOI: http://dx.doi.org/10.1007/978-3-540-85980-2_27.
8. Meinedo, H., Abad, A., Pellegrini, T., Neto, J., Trancoso, I.: The L2F Broadcast News Speech Recognition System. In: FALA 2010: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, Vigo, Spain, November 10-12, 2010. (2010)
9. Batista, F., Caseiro, D., Mamede, N., Trancoso, I.: Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news. Speech Communication **50** (2008) 847–862 DOI: http://dx.doi.org/10.1016/j.specom.2008.05.008.
10. Moreau, N., Hamon, O., Mostefa, D., Rosset, S., Galibert, O., Lamel, L., Turmo, J., Comas, P.R., Rosso, P., Buscaldi, D., Choukri, K.: Evaluation Protocol and Tools for Question-Answering on Speech Transcripts. In: Proceedings of the 7th Language Resources and Evaluation Conference - LREC 2010. (2010) 2769–2773
11. Carvalho, G., de Matos, D.M., Rocio, V.: Document Retrieval for Question Answering: A Quantitative Evaluation of Text Preprocessing. In: Proceedings of ACM first Ph.D. Workshop, PIKM 2007, in the 16th ACM Conference on Information and Knowledge Management, CIKM 2007, Lisboa, Portugal, November 5-10, 2007, ACM (2007) 125–130 ISBN: 978-1-59593-832-9 DOI: http://dx.doi.org/10.1145/1316874.1316894.