# IdSay: Question Answering for Portuguese

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@univ-ab.pt,vjr@univ-ab.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.** IdSay is an open domain Question Answering (QA) system for Portuguese. Its current version can be considered a baseline version, using mainly techniques from the area of Information Retrieval (IR). The only external information it uses besides the text collections is lexical information for Portuguese. It was submitted to the monolingual Portuguese task of the QA track of the Cross-Language Evaluation Forum 2008 (QA@CLEF) for the first time, and it answered correctly to 65 of the 200 questions in the first answer, and to 85 answers considering the three answers that could be returned per question. Generally, the types of questions that are answered better by IdSay system are measure factoids, count factoids and definitions, but there is still work to be done in these areas, as well as in the treatment of time. List questions, location and people/organization factoids are the types of question with more room for improvement.

## 1 Introduction

The objective of a 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.

IdSay (I'd Say or I dare Say) is an open domain QA system for Portuguese that was developed from scratch, with the objective of optimizing computational space and time, so that response could be fast. It was submitted to the monolingual Portuguese task of the QA track of the Cross-Language Evaluation Forum 2008 (QA@CLEF) for the first time. IdSay results placed it in third place among the other five systems that had participated in previous campaigns. Details of the task, and comparative results can be found in the overview of the QA track [1].

In Sect. 2 we describe IdSay briefly. In Sect. 3 we analyse the results obtained in QA@CLEF 2008, and in Sect. 4 we end with conclusions and future work.

C. Peters et al. (Eds.): CLEF 2008, LNCS 5706, pp. 345–352, 2009.

## 2 The IdSay System

Developing a QA system combines the task of treating large quantities of unstructured data (text), and the need to have a good understanding of the text to produce exact and short answers. Therefore it is natural that the areas of IR and natural language processing (NLP) are the foundations of these systems.

This is the approach we intend to follow in building IdSay system. We started by developing the core version of the system, which is based on information retrieval techniques. We chose this option for two main reasons: Firstly because we want to have a baseline to compare and draw conclusions of the effectiveness of the further NLP enhancements we plan to implement. Secondly because we intend to have an efficient retrieval base that can work as independently of the language as possible to reuse with different languages in the future.

The present version of IdSay 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]. In the rest of this section we briefly describe IdSay system, starting by the information indexing in Sect. 2.1, followed by an overview of the architecture of the system in Sect. 2.2.

### 2.1 Information Indexing

IdSay system is based on indexing techniques that were developed from scratch using C++. The IR engine was built with cross-language usage in mind, so we tried to develop it modularly, with the language-specific information clearly separated from generic components. For this purpose we analyse the input text data in successive levels, building an index file for each layer.

**Level 1 Document Level.** The documents are kept as close to the original text as possible, apart from the compression techniques used. It includes also tokenization and the minimal pre-processing to allow efficient retrieval, namely separation of words with spaces and lowercase conversion.

**Level 2 Lemmatization or Stemming.** According to the results of our previous work [3], in which lemmatization and stemming were compared, we opted for doing only lemmatization[1]. We intend however, in future versions of the system, to try different stemming techniques and lemmatization using a different lexicon. We do not remove stop words from the texts. This level corresponds to making equivalence classes based on related words at a linguistic level, and therefore it is one of the levels that is more language-specific.

**Level 3 Entities.** At level 3, which we call the entity level, we find all sequences of words that co-occur often in the text collections, and if their number of occurrences is higher than a given threshold (100 seems to be a reasonable

[1] Both options are available; when we say we use lemmatization, we are talking about the system setup for QA@CLEF.

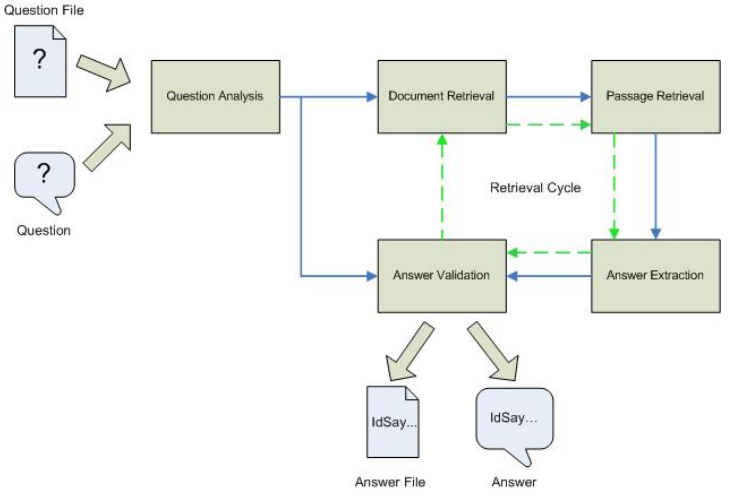


**Fig. 1.** IdSay system architecture

value), we consider them an entity whether it corresponds to a meaningful entity, like the name of an organization, or to a common string of words. For the time being, we rely on our ranking mechanism to eliminate the second kind of entities from answers, but we may do some further work in this area in the future.

### 2.2 System Overview

IdSay accepts either a question written by the user (manual interface), or a set of questions in an XML file (automatic interface). Each question is analysed in the question analysis module to determine the question type and other variables to be used in the answer extraction and validation modules. 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. 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 (number of words) up to a given limit (60). 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 (retrieval cycle).The global architecture of IdSay is presented in Fig. 1.

The index files for the text collection[2] occupy 1.15 GB of disk space, and took about 4 hours to build. The load time is around 1 minute, and the time

[2] The text collection occupies around 9 GB of disk space, in over 600,000 files. More details on the collection can be found in [1].

to process 200 questions is less than 1 minute. These values correspond to tests using a machine with an AMD Athlon 64 processor (2.21 GHz), with 4GB of RAM, running Windows XP.

## 3 QA@CLEF 2008 Results

In the present section we analyse the results obtained by IdSay. First we look into the evaluation metrics that describe the overall performance of the system, and proceed with a more detailed question based analysis.

### 3.1 Evaluation Metrics

The main evaluation metric used in QA@CLEF 2008 is accuracy over the first answer, which is the average of first answers that where judged to be correct. We also calculated the accuracy over all answers because it is also a common measure used for QA systems. Another metric used is MRR (Mean Reciprocal Rank) which is the mean of the reciprocal of the rank of the first answer that is correct for each question, as defined in [4]. Table 1. summarizes the results of IdSay system.

**Table 1.** IdSay results overview

| Accuracy over the first answer | Accuracy over all answers | MRR |
|---|---|---|
| 32.500% | 42.500% | 0.37083 |

### 3.2 Detailed Analysis of Results

IdSay has different approaches according to different criteria, for instance, specific procedures regarding question category and type. In the present section we analyse our results, covering different characteristics of the questions.

**Results by Question Category.** Three question categories are considered in QA@CLEF, namely *F* (factoids), *D* (definitions) and *L* (closed list questions).The results obtained by IdSay are summarized in Table 2.

**Table 2.** Results by category

| Question Category | Total Questions | Right | Wrong | ineXact | Unsupported | Accuracy |
|---|---|---|---|---|---|---|
| *F* | 162 | 47 | 100 | 7 | 8 | 29.012% |
| *D* | 28 | 18 | 10 | 0 | 0 | 64.286% |
| *L* | 10 | 0 | 9 | 1 | 0 | 0% |

The results show a stronger ability for the system to answer definition questions than factoids, which was expected due to the valuable aid of having an encyclopaedic data collection. The low value obtained for list questions is not a surprise because we did not have the time to treat this category of questions, so these are treated as factoids.

**Definition Questions.** This type of question generally occurs in the form: "O que é X?" [What is X?] or "Quem é X?" [Who is X?], in which we consider X the reference entity. IdSay starts by searching for the reference entity in Wikipedia, looking for a page for this concept. If such a page is found, the beginning of the page is returned as the answer.

The majority of definition questions were of the type "O que ser X?" [What to be X?][3]. IdSay answered correctly to half of them based on Wikipedia pages. If Wikipedia does not provide a definition, we follow the default procedure of searching the data collection in search for occurrences of the reference entity. An example of a correct definition found via the default procedure is (Question#66 O que é o jagertee?) [What is jagertee?], for which the answer was found within the data collection, in a sentence "o jagertee é chá com adição de rum" [jagertee is tee with addition of rum].

There were 7 definition questions of the type "Quem ser X?" [Who to be X?], of which IdSay answered 5 correctly based on Wikipedia pages. The two questions not answered correctly were (Question#23 Quem é FHC?) [Who is FHC?] and (Question#41 Quem é Narcís Serra?) [Who is Narcís Serra?]. The first corresponds to a Wikipedia page that is not found because the keyword FHC is not the name of the page for former Brazilian President Fernando Henrique Cardoso (but rather a redirect). In the second case, there is no Wikipedia page for Narcís Serra, and although in this case two news articles are found with the information, the answers were wrong due to extraction problems.

**Factoids Results by Question Type.** We consider the following types of questions: *P* - person/organization, *D* - date/time, *L* - location, *C* - count, *M* - measure, *O* - Other. We will start by analysing the results for the types for which we developed special procedures because they involved numeric values: *C*, *M* and *D*. We consider the assessment of the question to be the best answer, using the following priority: R, U, X and W[4].

Table 3 presents the results of IdSay for the type of factoids count, measure and date. We procced with an analysis of these results.

**Table 3.** Results by question type

| Question Type | # Questions | Right | Wrong | Unsupported | ineXact |
|---|---|---|---|---|---|
| Count | 19 | 13 (68.4%) | 5 (26.3%) | 1 (5.3%) | 0 (0%) |
| Measure | 12 | 9 (75.0%) | 2 (16.7%) | 1 (8.3%) | 0 (0%) |
| Date | 24 | 11 (45.8%) | 12 (50.0%) | 0 (0%) | 1 (4.2%) |

*Factoids Count.* These questions usually start by "Quantos/as X " [How many X]. X usually represents what we are tying to count. The general form of the

[3] We use the lemmatized form of the verb to cover the several tenses occuring in the questions.

[4] For example, if a question has three answers judged W, X and U we consider the U answer.

answer is usually a number followed by X. There were 20 count questions, with very diverse instances of X, namely esposas, faixas, províncias, repúblicas, actos, atletas, estados, filhos, filmes, gêneros, habitantes, jogadores, ossos, refugiados, votos [wives, stripes, provinces, republics, acts, athletes, states, sons, movies, gender, inhabitants, players, bones, refugees, votes].

An example of a correct answer is (Question#70 Quantas províncias tem a Ucrânia?) [How many provinces does Ukraine have?]. In the question, the reference entity Ukraine was identified and the identification of the unit allowed the correct answer to be found: 24 provinces. The case of (Question#10 Quantas províncias tem a Catalunha?) [How many provinces does Catalonia have?] is similar, with 51 documents retrieved that produced the answer "4 provinces" supported by more than one passage. However the answer was considered unsupported, due to the choice of the shortest passage. As an example of a question that produced wrong answers, we can look at (Question#18 Quantos ossos têm a face?[sic]) [How many bones do the face have?]. Although the question is incorrectly formulated (agreement is violated because the verb should be singular), the lemmatization took care of that and produced the search string "bone to have face". However, the answers produced were incorrect (number of bones of parts of the face, as the nose, returned) because the correct answer occurred in a phrase using the construction "é constituída por" [consists of] instead of the verb "ter" [to have].

*Factoids Measure.* This type of question is similar to the previous one, and generally occurs if the form of "Qual/ais .. o/a X de " [What the X of ] in which X is a measure, which can have several units. The answer is generally a numerical value in the correct units for the measure. There were several cases of measures in the question set: altura, área, dotação, envergadura, largura, temperatura, comprimento [height, area, money value, bulkiness, width, temperature, length]. IdSay supports several systems of measures and the corresponding units implemented in the manner of authority lists as described in [5]. It allows the search of the answers of the correct type.

An example of a correct answer is (Question#142 Qual é a área da Groenlândia?) [What is the area of Greenland?], for which only the value of the area "2 170 600 km 2" is returned and in the same passage there are other numbers, that would also be returned if we did not check the area units. The incorrect answers were given for questions that supposedly should produce NIL answers.

*Factoids Date.* The most common form of occurrence for this type of question is in questions starting by "Quando" [When], though there are also 4 questions staring by "Em que ano" [In which year]. IdSay has a specific treatment of dates, starting with the pre-processing of the texts, and also in the extraction of the answer. However this treatment is not fully developed, for instance the temporal restrictions are not taken into account. Therefore, the results achieved for this type are worse than for the preceding two types. The low accuracy for temporally restricted questions, 18.750%, can also be interpreted in light of this limitation.

An example of a correct answer is (Question#86 Quando é que ele tomou posse?) [When was he empowered?], which is also an example of a question that belongs to a cluster with first question (Question#85 Quantos votos teve o Lula nas eleições presidenciais de 2002?) [How many votes had Lula in the presidential election of 2002?]. Although Question#85 was not successfully answered, the reference to Lula (Brazilian President Luiz Inácio Lula da Silva) is correctly resolved in Question#86 (reference resolution based on the question, not the answer). As for the 12 wrong answers there are several aspects that contribute to that, there are questions about periods that were not treated by the system, and there is a need to treat date information from Wikipedia in a more practical way, e.g. the listed items in such pages are not terminated, so events tend to be mixed up in the resulting text.

*Factoids Person.* This type of question generally appears in a form starting by "Quem" [Who], but that is not always the case. The results for this type had an overall accuracy of 34%, which is in line with the general performance of the system. Examples of correct answers were (Question#92 Quem fundou a escola estóica?) [Who founded the stoic school?] (Question#143 Quem foi a primeira mulher no espaço?) [Who was the first woman in the space?] for which the system gives the correct answers (Zenão de Cítio and Valentina Tershkova, respectively) but they are accompanied by wrong second and third answers, that have different information related to the subject. We must therefore find a way to filter entities of type person. As stated in Sect. 2, IdSay keeps two separate indexes for words and for entities (two words or more). In the case of these two questions, the number of documents retrieved searching only for words were 11 for Question#92 and 1991 for Question#143. After combining the search for entities the number of documents decreased to 2 and 75, respectively. The case of Question#143 clearly shows an example of the utility in combining the search by single word with the search for entities.

**NIL Accuracy.** About the NIL accuracy, the reported value of 16.667% (2 right answers out of 12) for IdSay indicates the need of improvement in our mechanism to determine how well a passage supports the answers, to minimize the negative effect of the retrieval cycle in relaxing constraints. However comparatively to the other systems IdSay has the highest performance in NIL accuracy.

## 4 Conclusions and Future Improvements

We found the results of our first participation at QA@CLEF very encouraging. The fact that these results were obtained with particularly challenging rules (even for veteran participants) seems to reinforce the validity of our approach.

The analysis of our participation at QA@CLEF shows that the retrieval component works reasonably well, but the answer extraction mechanism is less efficient and is generally responsible for the wrong answers produced by the system. We expect that the introduction of NLP techniques will help in this regard. Another area that we identified that can benefit from these techniques is the answer

validation module. In this module the ranking of answers by frequency means that we produce the answer that appears most frequently in the passages extracted from the data collection. This means that an answer may be supported by several passages, but we can only give one as support. In this participation we chose the shortest one, but in several cases this option led for the support to be considered unsatisfactory by the assessors. We must therefore introduce an analysis of the passages to determine how strongly they support the answer.

Regarding the setup of the system, we find lemmatization a good choice as a whole, since it provides an efficient search, with just one case of a definition being wrong on its account.

As for short term improvements, these include attributing a confidence score to each answer, treating temporally restricted questions and the improvement of co-references between questions. The scoring mechanism of the answers is already partially implemented, since several supports for an answer are already considered, with different weights attributed to different kinds of occurrences.

As for future enhancements, besides of the introduction of NLP methods, we intend to accommodate semantic relations between concepts by adding further levels of indexing. As an example, we would like to introduce equivalences at a conceptual level, for instance by means of a thesaurus. Another future direction we intend to follow is introducing other languages besides Portuguese.

## References

[1] Forner, P., et al.: Overview of the CLEF 2008 Multilingual Question Answering Track. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 262–295. Springer, Heidelberg (2009)

[2] Alves, M.A.: Engenharia do Léxico Computacional: princípios, tecnologia e o caso das palavras compostas. Mestrado em Engenharia Informática Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (2002)

[3] Carvalho, G., Martins de Matos, D., Rocio, V.: Document retrieval for question answering: a quantitative evaluation of text preprocessing. In: Proceedings of the ACM first Ph.D. workshop in CIKM (ACM), pp. 125–130 (2007)

[4] Magnini, B., et al.: The Multiple Language Question Answering Track at CLEF 2003. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 471–486. Springer, Heidelberg (2004)

[5] Prager, J.: Open-Domain Question-Answering. Foundations and Trends® in Information Retrieval (Now Publishers) 1(2), 91–231 (2006)