IdSay: Question Answering for Portuguese

Gracinda Carvalho
Universidade Aberta
L2F/INESC-ID Lisboa
CITI – FCT/UNL
Rua da Escola Politécnica, 147
1269-001 Lisboa, Portugal
gracindac@univ-ab.pt

David Martins de Matos
L2F/INESC-ID Lisboa
Instituto Superior
Técnico/UTL
Rua Alves Redol 9
1000-029 Lisboa, Portugal
david.matos@inesc-id.pt

Vitor Rocio
Universidade Aberta
CITI – FCT/UNL
Rua da Escola Politécnica, 147
1269-001 Lisboa, Portugal
vjr@univ-ab.pt

Abstract

IdSay is an open domain Question Answering system for Portuguese that was developed from scratch. Its current version can be considered a baseline version, using mainly techniques from the area of Information Retrieval. The only external information that it uses besides the text collections is lexical information for Portuguese.

It was submitted to the monolingual Portuguese task of the Question Answering track of the Cross-Language Evaluation Forum 2008 (QA@CLEF) for the first time, in which 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 evolution.

Categories and Subject Descriptors:
H.3 [**Information Systems-Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing, H.3.2 Information Storage, H.3.3 Information Search and Retrieval, H.3.4 Systems and Software
I.2 [**Computing Methodologies-Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms:
Experimentation, Measurement, Performance, Languages

Keywords:
Information Retrieval, Question Answering

1 Introduction

The objective of a Question Answering 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, or the blogosphere, or even the World Wide Web, there is an increasing interest in this type of system.

IdSay (I'd Say or I dare Say) is an open domain Question Answering system for Portuguese that was developed from scratch, with the objective of optimizing resources, so that response time could be short. Its current version can be considered a baseline version, using mainly techniques from the area of Information Retrieval.

It was submitted to the monolingual Portuguese task of the Question Answering track of the Cross-Language Evaluation Forum 2008 (QA@CLEF) for the first time.

We end this section with the description of the task, and in section 2 we make a brief description of IdSay system. In section 3 we analyse the results obtained in QA@CLEF 2008, and in section 4 and 5 we end with some conclusions and future work respectively.

1.1 QA@CLEF: Main Task Description

The Portuguese monolingual task of QA@CLEF consists in finding answers to 200 questions within a text collection, all in Portuguese.

The text collection used consists of newspaper articles from Portuguese newspaper “Público” and from Brazilian newspaper “Folha de São Paulo” from the years 1994 and 1995, and a frozen version of the HTML edition of the Portuguese Wikipedia[1] from November 2006.

Systems should be able to answer three categories of questions, namely factoids[2], definitions and closed list questions, some of which may include a temporal restriction. The type of the questions can very diverse, ranging from questions about persons, organizations or locations, for instance. Neither the question category, nor its type or if it is temporally restricted are explicitly given with the question, therefore it is up to the system to find out that information. The questions can be stand-alone questions, or organized in clusters of up to four questions about a common topic, with the topic being determined by the first question/answer pair. Questions 2 to 4 of the cluster can contain co-references to the cluster’s topic. The clusters are identified by a number that is given to the system. It is possible that a question has no answer in the data collection, in which case the answer should be NIL. Each question, except NIL questions, must be supported with an excerpt from a document in the collection, therefore an answer consists of the exact answer (with no more information than strictly needed), an identifier from a document in the text collection, and an excerpt of up to 700 characters from that document that supports the answer[3]. The system may produce up to three answers for each question, and two complete answer files can be submitted for evaluation.

2 IdSay System

In this section we briefly describe our system, starting by the information indexing in subsection 2.1, followed by a subsection for each module of the system.

Developing a Question Answering system can be interpreted as an engineering problem of combining several types of components and techniques to achieve, in an efficient way, the goal stated in the beginning of the introduction. Given the nature of 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, it is natural that the areas of Information Retrieval and Natural Language Processing 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, that is based in Information Retrieval techniques. We made 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 intend 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 therefore as close as possible to simple keyword search. The only external information that we use besides the text collections is lexical information for Portuguese **[1]**.

The global architecture of IdSay is presented in Fig. 1.

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 user can select if he wants to search for the information in the entire data collection, or in one or both the newspaper collections or just Wikipedia. It can also specify the maximum number of answers allowed, and other system related parameters, for example if lemmatization or stemming should be considered, or the detail of information to be shown by the system. The system allows for

[1] http://pt.wikipedia.org/

[2] We use in the text the term “factoid”, because it is widely used in the literature. However we would prefer the term “questions about facts”, or “factual questions”, because of the meaning of “factoid” as “an invented fact usually taken as true”.

[3] In fact, up to three separate excerpts from the same document can be provided, as long as the limit in characters is not exceeded.

various statistics of the data to be consulted. In the manual interface, the system is not prepared to treat co-reference between questions.

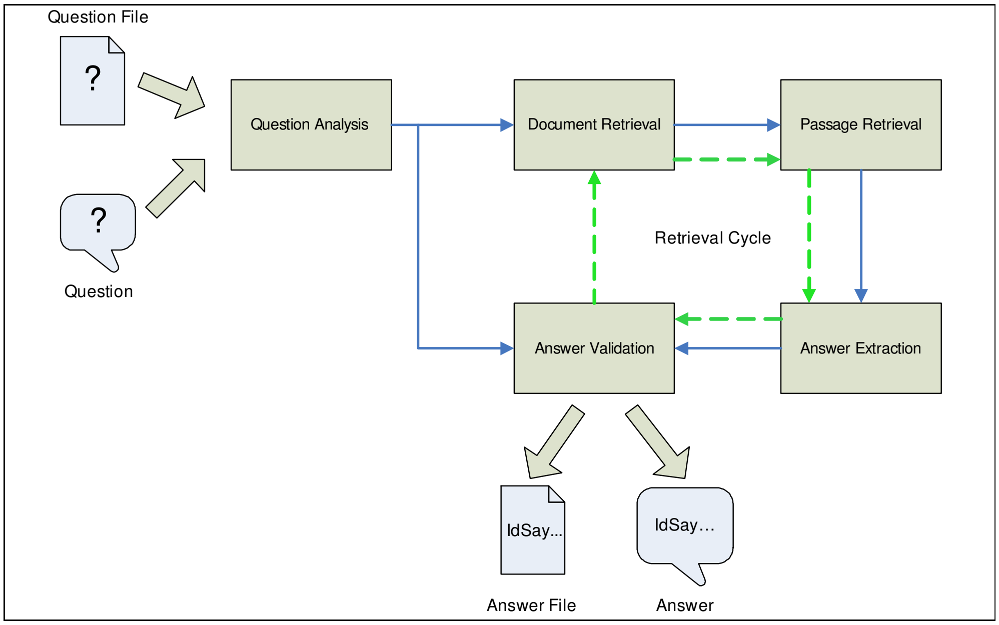

Fig. 1 – IdSay system architecture

Each question is analysed in the question analysis module (section 2.2) 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 (section 2.3) to produce a list of documents that match both. This list of documents is then processed by the passage retrieval module (section 2.4), 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 (section 2.5), where short segments of text (candidate answers) are produced, that are then passed on to the answer validation module (section 2.6). 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.

2.1 Information Indexing

IdSay's architecture is based on indexing techniques that were developed from scratch for the system. However, these techniques are general purpose IR and are not specific for Question Answering alone. The IR engine was also built with cross-language usage in mind, so we tried to develop it modularly, with the language specific information clearly separated from other generic components. For this purpose we analyse the input text data in successive levels, building an index file for each layer:

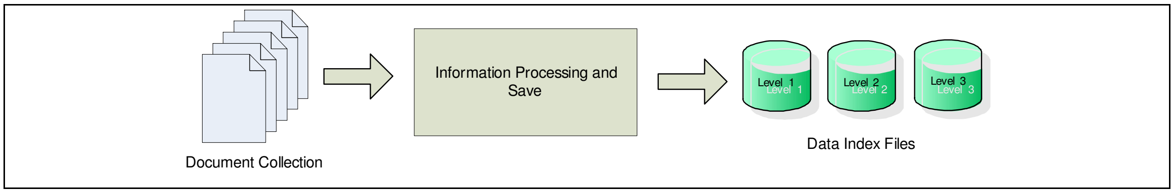

Fig. 2 –Information Indexing in IdSay

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 [2] in which lemmatization and stemming were compared we opted for doing only lemmatization[4]. 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 form the texts, but words maybe removed at a later stage, during the retrieval cycle.

Since this process is used to increase the retrieval efficiency in finding relevant documents, and having in mind the fact that we may need the original information (or as close as what we can have, that is to say level 1 index) we store more information in the level 2, but we also keep the information of level 1. That prevents us from having to go back to the original document, which helps us improve the performance in terms of response time.

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 markedly language-specific.

Level 3 – Entities

At level 3, which we call the entity level, we find all sequences of words that occur often in the text collections, and if they occur more than a given threshold, 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, but we may do some further work in this area in the future.

The index files for the text collection occupies 1.15 GB of disk space, and took about 4 hours to build. The load time is around 1 minute, and the time to process 200 questions is less than 1 minute.

2.2 Question Analysis

This module processes the question and produces a search string. It needs to identify the category of the question (within the CLEF standard categories factoid, definition, and closed list question) and the type of answer expected. This analysis depends on finding specific patterns for Portuguese, which are normally used to formulate each kind of question. For instance, if we have a question in the form “O que é X?” [What is X?] or “Quem é X?” [Who is X?], we conclude that the category of the question is definition[5].

If we are unable to determine the question category and type, we treat it as a generic question, for which we have a default procedure.

Besides the category and type of the question, which will be used at a later stage to get answers of the correct type, this module also identifies the appropriate information that should be used to guide the search for documents related to the question, which we call reference entities. These entities are also used for co-reference resolution in the case of clusters of questions.

The question is searched for reference entities in the following manner: in a first stage, we rely on the indications of the user. Therefore if there are words capitalized, or words enclosed in single or double quotation marks or guillemets, we assume them to be entities[6]. In a second stage, the text of the question is searched looking for entities that we have found in the text collections, and these are registered to be used in the search.

[4] The two options are available, when we say we use lemmatization, we are talking of the system setup chosen for the QA@CLEF evaluation.

[5] Since we use lemmatization, the pattern we look for is in fact “Quem ser X?” [Who be X?], which covers several tenses that can be used in the question, like “Quem foi X?” [Who was X?] as well as the present tense.

[6] If the several contiguous words are all capitalized, or included in brackets, we consider them to be just an entity.

2.3 Document Retrieval

This module takes the words from the search string produced previously and generates a list of documents that contains all of the words in the search string, and also all the entities.

This is done efficiently, since we have the documents indexed by words and by entities. The document list is built through the intersection of both the list of documents that contain each single word and the list of documents that contain the entities present in the question.

If the list of documents is empty, the process is repeated, removing the most frequent word from the search string. This process is identified in Fig. 1 as retrieval cycle, and intends to increase the possibility of finding the correct answer, turning down the words with higher frequency of occurrence and therefore with lower discriminative power.

The only exception for this rule is when we are looking for the definition of a concept that has a page in Wikipedia. In this case, the answer of IdSay corresponds to the first sentence in the page.

2.4 Passage Retrieval

The aim of this module is, given a list of documents, to produce all the passages from the documents where the words we are looking for occur. The passage length should not exceed a given limit (currently, 60 words).

This is done in the following way: each document is searched once for the words of the question. Each time a word is found, its position in the document is registered. After storing this information for a newly found word, we check if all the words already have a position registered. If that is the case we check the total length of the passage by subtracting to the current position the minimum of the set of positions of all words. If the length does not exceed the limit we add this passage to the passage list.

Each passage is then adjusted, adding words to the beginning and to the end, in such a way that the passage corresponds, as much as possible, to one or more full sentences. For this purpose two punctuation marks sets are defined: the *terminators* which include for instance full stop (.), question mark (?) and exclamation mark (!) and the *separators* that include for instance the comma (,), semi-colon (;) and some words as 'e' [and] 'ou'[or].

The adjustment is made until the nearest *terminator* is found, both before the passage and after the passage. If a terminator in not found within a distance in words that allows the passage length limit to be respected, then the system searches for the nearest *separator*, and if it also exceeds the length limit then the passage of the maximum length is considered, even if it breaks sentences in the middle.

Each document in the list is searched, with the corresponding passages (which can be zero for a given document) being added to a global passage list common to all documents.

2.5 Answer Extraction

The input for this component of the system are passages, and from them we intend to extract answers, that is to say, we intend to eliminate the unnecessary portion of the passage, retaining no more information than what is absolutely needed to answer the question.

We analyse each passage searching for entities. Each entity found is considered a candidate answer. If no entities are found, the low frequency words are considered candidate answers.

A word is considered low frequency if its frequency is less than the double of the frequency of the least frequent word in the passage. All frequencies considered are the absolute frequencies in all the text collection.

If the question category is *D* (definition) the candidate answers considered are the phrases immediately before and after the word or entity for which the definition is being sought.

The extraction phase takes into account the category and type of the question if they were identified in the question analysis phase, in a way that prevents us from extracting answers that are not related to what we are looking for. For instance, if the answer type is date, time or numeric (count or measure) then the system searches the passage looking for numeric values followed by the corresponding units if they occur in the passage.

2.6 Answer Validation

This module validates if the answers produced are the best ones given the other answers and their supporting passages. It receives the answers with the respective passages from where the answers were extracted, and orders the answers by the scoring based on frequency.

The answers that have the same supporting document, even if the passages are slightly different, are merged if they are close in the document. For instance, if the passage "… Ms. X is the prime-minister of country Y…" supports two different answers, "prime-minister" and "Y", a new answer is produced joining both answers from the passage: "prime-minister of country Y".

Afterwards, the list of answers is filtered, by removing the answers with common words, with lower score. For instance, using the same example as above, if the answer list contains, in a lower position, an answer "minister", it will be removed since is contained in "prime-minister of country Y".

Finally, the support used for each question is the smallest passage associated with each answer.

3 QA@CLEF 2008 Evaluation Campaign Results

In the present chapter we analyse the results obtained by IdSay system. 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 answers returned by the system are evaluated manually by human assessors. The metrics used to rank the systems are based on a Boolean function that takes into account the judgment of the assessors. The definition of that function, which we will name F_{ij} for the case of answer *j* to question *i* is $F_{ij} = 1$ in case answer *j* to question *i* is considered correct ('R'), or $F_{ij} = 0$ otherwise (i.e. in case answer *j* to question *i* is considered wrong, unsupported or inexact ('W','U','X'), or in case the system does not provide an answer *j* to question *i*).

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.

Considering a set consisting of Q questions, and if for each question up to N answers may be provided, a more formal definition can be given as:

$$Accuracy\ over\ the\ first\ answer = \frac{1}{Q}\sum_{i=1}^{Q} F_{i1}$$

In QA@CLEF, the values are Q=200 and N=3.

We also calculated the accuracy over all answers because it is also a common measure used for Question Answering systems. If we define R_i, for question *i*, as:

$$R_i = \begin{cases} 1 & if\ there\ is\ j = 1..N\ for\ which\ F_{ij} = 1 \\ 0 & if\ F_{ij} = 0\ for\ all\ j = 1..N \end{cases}$$

We can define the accuracy over all answers in the following way:

$$Accuracy\ over\ all\ answers = \frac{1}{Q}\sum_{i=1}^{Q} R_i$$

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.

If we define J_i, for question *i,* as:

$$J_i = \begin{cases} \dfrac{1}{\min\limits_{j=1..N} \{ j : F_{ij} = 1 \}} & \textit{if there is } j = 1..N \textit{ for which } F_{ij} = 1 \\ 0 & \textit{if } F_{ij} = 0 \textit{ for all } j = 1..N \end{cases}$$

The definition of MRR would be:

$$MRR = \frac{1}{Q} \sum_{i=1}^{Q} J_i$$

This metric takes into account the correct answers returned for each question regardless of the fact that they where returned in the first position or not. Correct answers that are given earlier in the answer list have a higher contribution to the score. For a given system the MRR should always produce an equal or higher score than the accuracy over the first answer score. All questions in the question set are considered, according to the definitions that are presented in the guidelines [4], and in the literature, for instance the overview of the first overview article of QA@CLEF [3].
However, the metric that was published as MRR for the systems at QA@CLEF2008 was slightly different, because it took into account only the answers for which multiple answers had been provided. This has several drawbacks, in our opinion, since it is trying to compare systems based on a different number of questions and besides it has the effect of penalizing systems that gave single correct answers. So if a system is confident enough in an answer not to provide more possible answers, and if the answer is correct, it gets worse results that if it choose to provide the same answer three times, or a second constant answer, like NIL. In the case of our system, it provided 34 single answers that were judged correct. In the limit cases, the result of this way of computing the metric are: (1) the value would always be 0 for a system that provided both single answers and multiple answers, with the correct answers being all given in the form of single answers, no matter the proportion between correct and incorrect answers, (2) the value is undefined for systems that provide single answers for all questions, regardless of their correctness.

We include in Table 1 the MRR for our system calculated according to the formula above because we consider it a good reference for comparison between systems.

Another metric that was provided for the systems was CWS (Confidence Weighted Score). It was used for the first time at the Question Answering track of TREC 2002 [6], in which a file was provided by the participants with just one answer per question, with the answers ordered not by question number, but in order of decreasing confidence in the answer.

A formal definition would be:

$$CWS = \frac{1}{Q} \sum_{i=1}^{Q} \frac{\sum_{j=1}^{i} F_{j1}}{i}$$

It is important to keep in mind that index *i* follows the answers in order from most confident response to least confidence response, and that there is exactly one answer per question (hence the index 1 for answer number in F_{j1}).

Since our system did not provide a confidence score for the questions (i.e. provided a score of 0.0 for all answers, as stated in the guidelines) we would expect CWS not to be calculated in this situation. The reason for that is that since we do not provide scores for the answers, there is no valid criterion to sort the answers, so it seems artificial to attribute an order to 200 equal values.

However, a score was attributed to the system, which took into account the first answer to each question, and using question number order. Moreover, a different version of the formula above was used, that we identify as CWS*, in which the contribution of earlier correct answers where not taken into account, if the current answer was not correct, which therefore yields lower scores.

$$CWS^* = \frac{1}{Q} \sum_{i=1}^{Q} \frac{F_{i1} \sum_{j=1}^{i} F_{j1}}{i}$$

We will make no further considerations on the values for CWS of system Idsay. We would like however to comment on the possible arbitrariness of CWS as a measure of the scoring capacities of a system: if the system provides the same score to several answers, which order should be chosen for these answers? Question number order (if it exists…)? Alphabetical order? Random order? Whatever the option taken, if some questions are correct and others are not, it can result in different values of CWS. Our system was just the extreme case of this situation, with all the answers with the same confidence score.

Table 1 summarizes the results of IdSay system.

Accuracy over the first answer **# %**	**Accuracy over all answers** **# %**	**MRR over multiple answers**	**MRR**	**CWS***	**CWS**
65 32.500%	85 42.500%	28.487%	37.083%	11.620%	34.447%

Table 1 – IdSay results overview

3.2 IdSay Results Detailed Analysis

As described above, 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 categories are considered in QA@CLEF, namely *F* questions (factoids), *D* questions (definitions) and *L* questions (closed list questions).

These values are provided by the Organizers, according to its classification in the three categories above[7]. The results obtained by IdSay are summarized in Table 2.

			Judgement				
		Total questions	Right	Wrong	ineXact	Unsupported	accuracy
Question Category	*F*	162	47	100	7	8	29.012%
	D	28	18	10	0	0	64.286%
	L	10	0	9	1	0	0%

Table 2 – Results by category

[7] We do not know the classification of the questions by category of the organization, so in the rest of the text we use our classification, therefore the values can be slightly different.

These results show a stronger ability for the system to answer definition questions than factoids, which was expected due to the valuable aid of the having an encyclopaedic data collection, since its aim is precisely to present definitions. The 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. In the case of the answer assessed as inexact, the list of three possibilities was given as the three answers to the question (Question#154 Por que estados corre o Havel?) [For which states does the Havel run?].

We start by making a more detailed analysis of definition questions, for which the type of question is not used and proceed with an analysis by question type for 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.

There were 22 definitions of the type "O que ser X?" [What to be X?], of which IdSay answered 11 correctly based in Wikipedia pages. There are a few cases in which the reference entity has an entry in Wikipedia, but there are several ways of referencing the same entity and therefore we do not get to the right page straight away. That is, for instance, the case of (Question#35 O que era o A6M Zero?) [What was the A6M Zero?]. However, we got to Wikipedia web page via retrieval of content, and manage to produce the correct answer, in this case but not in others. There is an advantage in dealing with the different ways of naming the same entity in Wikipedia, and we did not do that for the current version of the system only because of lack of time. A similar problem happens in (Question#127 O que são os forcados?) [What are forcados?] which in Wikipedia is identified by the singular form and not the plural of the concept. We tryed to get the definition by lemmatization, however the word retrieved is the lemma of the verb, and unluckily there is a town in Spain that corresponds to a form of that verb, and that is the definition returned. In this case, we would have to consider besides of the alternative ways of naming the same entity in Wikipedia, the plural/singular variation. 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?]. There is only one occurrence of this entity in the data collection, in a sentence "o jaggertee é chá com adição de rum" [jaggertee is tee with addition of rum], which allows the system to retrieve the correct answer "chá com adição de rum" [tee with addition of rum]. However many definitions are not found in these circumstances: that is the case of (Question#80 O que é o IPM em Portugal?) [What is the IPM in Portugal?] for which the acronym of this institution is found many times together with other institution names, which does not allow for the correct answer to be retrieved.

There were 7 definition questions of the type "Quem é X?" [Who is X?], of which IdSay answered 5 correctly based on Wikipedia pages. The two questions that were wrong (Question#23 Quem é FHC?) [Who is FHC?] and (Question#41 Quem é Narcís Serra?) [Who is Narcis Serra?] the first corresponding to a Wikipedia page that is not found (again) because the keyword FHC is not the name of the page for former Brazilian President Fernando Henrique Cardoso (but rather a redirect), and the second case because there is no Wikipedia page and although in this case two news articles are found with the information on Narcis Serra, the answers are 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*. The values we present are based on our manual classification of the questions in the types above. Each answer of each question is assessed as 'R', 'U', 'X' or 'W', so we consider the assessment of the question to be the best answer, using the following priority: 'R', 'U', 'X' and 'W'[8]. For each type, we try to analyse an example to draw conclusions about the strength and weakness of the system in that type, except for type other, because the questions can be so different from one another that an almost per question analysis would be required.

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

Factoids – Count

These questions usually start by “Quantos/as X …” [How many X…]. X usually represents what we are tying to count, followed by the rest of the phrase. The general form of the question 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, habitants, jogadores, ossos, refugiados, votos [wives, stripes, provinces, republics, acts, athletes, states, sons, movies, gender, inhabitants, players, bones, refugees, votes].
The results of IdSay for this type of factoids are presented in Table 3.

# Questions	Right		Wrong		Unsupported		ineXact	
	#	%	#	%	#	%	#	%
19	13	68.4	5	26.3	1	5.3	0	0

Table 3 – Results for Count Questions

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 to look for was provinces. The search string was in this case “província ter a ucrânia” [provinces to have ukraine] for which 71 documents were retrieved, where the correct answer was 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 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 in terms of concordance in number of the verb, which is in the plural instead of singular as it should, 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 a different verb instead of tem [has] the construction “é constituída por” [consists of].

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 the authority lists described in [5]. It allows the search of the answers of the correct type. The addition of new measures is an easy process; however, the implemented lists should be stable. The results of IdSay for this type of factoids are presented in Table 4.

# Questions	Right		Wrong		Unsupported		ineXact	
	#	%	#	%	#	%	#	%
12	9	75.0	2	16.7	1	8.3	0	0

Table 4 - Results for Measure Questions

An example of a correct answer is (Question#142 Qual é a área da Groenlândia?)[What is the area of Greenland?]. For area measure we have cm 2, m 2, km 2, hectar, hectares, ha. The occurrences of “km2”, “km^2” or other forms commonly used are normalized to “km 2” in the pre-processing of the data collection occurring at level1. For this question, 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 normalization of the unit measures and numerical values is part of the pre-processing of the data collection occurring at level 1. The incorrect answers were given for questions that supposedly should produce NIL answers. The unsupported answer is (Question#163 Qual o comprimento da Ponte do Øresund?)[What is the length of the bridge of Øresund?] for which the support that was taken from a news article was considered insufficient.

Factoids – Date

The most common form of occurrence for this type of question is in questions starting by “Quando…” [When…], tough there are also 4 question 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 that for the preceding two types. The low percentage of accuracy in temporally restricted questions, 18.750%, can also be interpreted in light of this limitation.

The results of IdSay for this type of factoids are presented in Table 5.

# Questions	Right		Wrong		Unsupported		ineXact	
	#	%	#	%	#	%	#	%
24	11	45.8	12	50.0	0	0	1	4.2

Table 5 - Results for Date Questions

As an example of a correct answer is (Question#86 Quando é que ele tomou posse?)[When was he empowered?], and is 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). The system produces the correct answer twice, the first answer is just the year, 2003, which is more frequent, but the support was not accepted because it said something like “President from 2003 to actuality”, however the second answer, that indicated January of 2003, was considered right. 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 the need to treat date information from Wikipedia in a more practical way, for instance, the listed items in such pages are not terminated, so events tend to be mixed up in the resulting text. Since we had no notion of the dates, each numeric value can be a potential date, also with qualifiers like a.c., which increases the difficulty to answer this kind of question, namely the decision that the answer should be NIL. An example of a wrong answer is (Question#17 Em que ano é que Ernie Els venceu o Dubai Open?)[In which year did Ernie Els win the Dubai Open?] for which the system produced several incorrect numeric values, tentatively, from the only news article found with information on the subject, but in which the year of 1994 could only be produced if one took into account the date of the news article itself and not its text.

Factoids – Person

This type of question generally appears in a form starting by “Quem” [Who], but that is not always the case. There are 37 questions in total starting by “Quem” [Who], of which 7 are definitions and were already covered in the corresponding section, and 5 occurring with different forms. The results for this type had an overall accuracy of 34%, which is according to 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, respectively Zenão de Cítio e Valentina Tershkova, 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. The search strings were respectively “fundar escola estóico” [to found school stoic] and “primeiro marido[9] no espaço” [first husband in space] and the number of documents retrieved were respectively 11/2 and 1991/75 before entities/after entities. The case of Question#143 shows an example of the utility in combining the search by single word with the search for entities (see section 2.3).

An example of a question that produced only wrong answers was (Question#160 Quem realizou «Os Pássaros»?) [Who directed «The Birds»?]. Since the title of the film is composed of words that occur very often in other contexts, other documents were retrieved that had no relation to the subject of the question. The search string was in this case “realizar o pássaro” [to direct the bird] and the number of documents retrieved was 317 (no entities were found in the question text).

[9] The lemma of mulher [woman] in our system is marido [husband].

A last example of questions of this type is (Question#15 Quem escreveu Fernão Capelo Gaivota?) [Who wrote Jonathan Livingston Seagull?] in which there were 6 documents each with several passages supporting the correct answer of “Richard Bach”, but since the shortest passage was chosen as support, it was assessed unsupported.

Factoids – Location

These questions generally begin by “Onde” [Where]. IdSay does not make a special treatment for location questions. The results for this kind of question are close to the overall performance of the system. The strong incidence of this type of question in clusters also contributes to raise its difficulty. We believe that better results can be achieved through a better treatment of this kind of information from Wikipedia.

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 effect of the retrieval cycle in relaxing constraints. We cannot however make a very accurate analysis because the 12 questions to which the NIL answer should be returned have not been divulged, and also there are some questions whose formulation makes it unclear if they should be “corrected” or a NIL answer is expected. An example of such questions is (Question#152 Qual é a capital de Dublin?) [What is the capital of Dublin?]. We considered that these questions should not be intentional, because we are in the context of an evaluation campaign, and the aim is to compare results form several systems. For this goal to be achieved there should be as little ambiguity as possible. If there is the interest in finding out how systems react to strange situations it could be done at the CLEF Workshop interactively, using frozen versions of the system.

4 Conclusions

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 evolution.

The ordering (ranking) of answers by frequency means that we produce the answer that appears most frequently in the context in the data collection. However, this procedure may lead to difficulties in the support of the answer, because the same answer can occur several times in the collections, with different supports. Since in the CLEF evaluation campaign only one support is allowed, we chose the shortest one found by the system. This may lead, and it did in many cases, for the support not to be the best possible (and in some cases answers were considered unsupported by the assessor). We will consider in the future the analysis of all passages (or at least more than one) to determine how strongly it supports the answer. It is the cases mentioned in the factoid-count analysis, Question#10, in the factoid-measure analysis, Question#163, and in the factoid-person analysis, Question#127.

Although the lemmatization process produces some strange results in the search strings, it provides an efficient search, with the search phase of the process generally finding the related documents. We find the lemmatization a good choice as a whole, with just one case of a definition being wrong on its account, as in the case of Question#127 already mentioned. The extraction part is less efficient and is generally responsible for the wrong answers produced by the system.

If we had a mechanism for synonyms for instance for verbs, questions like that mentioned in the factoid-count analysis Question#18, would be answered successfully.

5 Future Improvements

As for short term improvements, these include attributing a confidence score to each question, treating temporally restricted questions and the improvement of co-references between questions. The scoring mechanism is already partially implemented, since several supports for an answer are already considered, and not only is the total number of occurrences taken into account, but also different weights are attributed to different kinds of occurrence. Also the fact of an answer being given in the form of a single answer reflects a higher confidence in the answer. What is missing is translating all this information into a single numeric value.

The treatment of closed list questions must also to be implemented, as well as a refinement on the treatment of co-references.

As for future enhancements, we would like to introduce equivalences at a conceptual level, for instance by means of a thesaurus. We intend to accommodate semantic relations between concepts by adding further levels of indexing.

Another future direction we intend to follow is introducing other languages besides Portuguese.

6 Acknowledgements

We thank the support of FCT[10] (Fundação para a Ciência e Tecnologia) grant FACC 08/5/1012. We thank Linguateca[11] for providing access to information from previous campaigns.

References

[1] **Alves, M. A. (2002) "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,.
http://www.liacc.up.pt/~maa/elc.html

[2] Carvalho, Gracinda; Martins de Matos, David; Rocio, Vitor (2007) "Document retrieval for question answering: a quantitative evaluation of text preprocessing".
In *Proceedings of the ACM first Ph.D. workshop in CIKM* (ACM), Pg. 125-130.
http://portal.acm.org/citation.cfm?id=1316874.1316894&coll=&dl=GUIDE&type=series&idx=SERIES772&part=series&WantType=Proceedings&title=CIKM&CFID=6916667&CFTOKEN=45826677

[3] Magnini, Bernardo et al. (2003) "The Multiple Language Question Answering Track at CLEF 2003".
In *Lecture Notes on Computer Science* (Springer), Vol. 3237/2004, Pg. 471-486.
http://www.springerlink.com/content/v4x1vxhjajp89hh5/

[4] "GUIDELINES for the PARTICIPANTS in QA@CLEF 2008".
Version-4/30/2008.
http://nlp.uned.es/clef-qa/QA@CLEF08_Guidelines-for-Participants.pdf

[5] Prager, John (2006) "Open-Domain Question–Answering".
In *Foundations and Trends® in Information Retrieval* (Now Publishers), Vol. 1: No 2, Pg. 91-231.
http://dx.doi.org/10.1561/1500000001

[6] Voorhees, Ellen M. (2003) "Overview of the TREC 2002 Question Answering Track".
In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*
http://trec.nist.gov/pubs/trec11/papers/QA11.pdf

[10] http://alfa.fct.mctes.pt/

[11] http://www.linguateca.pt/