Priberam's question answering system in QA@CLEF 2007

Carlos Amaral, Adán Cassan, Helena Figueira,
André Martins, Afonso Mendes, Pedro Mendes, Cláudia Pinto, Daniel Vidal

Priberam Informática
Alameda D. Afonso Henriques, 41 - 2.º Esq.
1000-123 Lisboa, Portugal
Tel.: +351 21 781 72 60
Fax: +351 21 781 72 79

{cma, ach, hgf, atm, amm, prm, cp, dpv}@priberam.pt

Abstract

This paper accounts for Priberam's participation in the monolingual question answering (QA) track of CLEF 2007. In previous participations, Priberam's QA system obtained encouraging results both in monolingual and cross-language tasks. This year we endowed the system with syntactical processing, in order to capture the syntactic structure of the question. The main goal was to obtain a more tuned question categorisation and consequently a more precise answer extraction. Besides this, we provided our system with the ability to handle topic-related questions and to use encyclopaedic sources like Wikipedia. The paper provides a description of the improvements made in the system, followed by the discussion of the results obtained in Portuguese and Spanish monolingual runs.

ACM Categories and Subject Descriptors
H.2 [Database Management]: H.2.3 Languages - Query Languages
H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms
Measurement, Performance, Experimentation, Languages.

Keywords
Question answering, Questions beyond factoids.

1 Introduction

Priberam has participated in the CLEF campaigns since 2005, where its QA system was evaluated in both monolingual [1, 2] and cross-language [3, 4] environments with rewarding results. In the framework of M-CAST[1], Priberam's system was also used for Portuguese, Polish and Czech and applied to a digital libraries project. For QA@CLEF 2007, we focused our participation on the Portuguese and Spanish monolingual tasks.

This year, the CLEF QA track presented two novelties with direct consequences in the evaluation of QA systems. First, the organisation introduced topic-related questions, that is, questions clustered around a common topic that might present anaphoric links between them. Second, it also added the open domain encyclopaedia Wikipedia as a target document collection to the already existent newspaper corpora.

[1] M-CAST – Multilingual Content Aggregation System based on TRUST Search Engine – was an European Commission co-financed project (EDC 22249 M-CAST), whose aim was the development of a multilingual infrastructure enabling content producers to access, search and integrate the assets of large multilingual text (and multimedia) collections (http://www.m-cast.infovide.pl).

Although its overall architecture was maintained, Priberam's QA system was subjected to several changes, both in the Portuguese and in the Spanish modules, the most relevant one being the introduction of syntactical question processing. These modifications, together with the introduction of the two simultaneous changes by the CLEF organisation, had an impact in the performance of our system, as we will show in the analysis of the results.

This paper is organised as follows: in section 2, we present the major adjustments made to the system, such as the work done in the syntactical processing of the question, and the necessary adaptations to deal with topic-related questions as well as with the Wikipedia encyclopaedic source; in section 3 we analyse and discuss the results of both monolingual runs; finally in section 4 we present the conclusions and future work.

2 Adaptations and improvements of the system

2.1 Question syntactic analysis

As reported in [2, 4], Priberam's QA system is based upon a five-step architecture: the indexing process, the question analysis, the document retrieval, the sentence retrieval, and the answer extraction. When a question is submitted and matches a given question pattern (QP), a category is assigned to it and a set of question answering patterns (QAPs) becomes active. Then, documents containing sentences with categories in common with the question (previously determined during indexation via answer patterns (APs)) are analysed; the active QAPs are then applied to each sentence in order to extract the possible answers. While the overall architecture remains unchanged, this year, following the conclusions taken from preceding evaluations, we implemented a mechanism for the syntactic treatment of questions.

In the former version of the QA system, the question analysis stage categorised questions according to a previously defined question typology, by matching a set of question patterns [2]. More than one category per question was allowed. Although a multicategorisation scheme has the advantage of allowing more than one category in cases where it is difficult to choose only one, the excess of categories was one of the causes for errors in the extraction of candidate answers in our former CLEF participations.

In the present version, and taking advantage of the company's linguistic technology developed for FLiP[2], QPs were enhanced with syntactical information. Currently, when a question is submitted, its syntactic structure is captured in a parsing stage that also determines the syntactic function of the question pivots. Parsing the question enables determining its main syntactic constituent, which we have called the *object*, as well as its secondary syntactic constituents, in case they exist[3]. For instance, in question 8 of the Portuguese test set "De que estado brasileiro foi governador Adhemar de Barros?" [Of which Brazilian state was Adhemar de Barros governor?], the object is *Adhemar de Barros* and the secondary constituents are *governador* and *estado brasileiro*. This information is then used to validate the category assigned to each question and to allow a more exact, syntactically based answer extraction. By differentiating pivots through syntactic tags, we also expect to decrease the amount of errors in the document retrieval stage, especially when it comes to long questions where there are too many pivots. In the answer extraction stage, we enhanced QAPs to include information about each pivot's syntactic specifications, which the system tries to match with the answer pivots. The syntactic information captured by parsing may also be useful for answer validation, an aspect that we intend to exploit in the future.

2.2 Handling topic-related questions

Priberam's QA system was adapted to handle topic-related questions. According to the QA@CLEF 2007 guidelines, these are clusters of questions related to the same topic and possibly containing anaphoric references between one question and the others. Although questions belonging to the same cluster are identified as such, no information is given about the topic, which has to be inferred from the first question/answer pair. A natural way to address this problem would consist of two steps: *(i)* detecting the topic, which requires deciding which named entity, event, object, or natural phenomena should be extracted from the first question/answer pair, and *(ii)* detecting the anaphora in each of the following questions and using the detected topic as the co-referent.

---

[2] *Ferramentas para a Língua Portuguesa*, Priberam's proofing tools package for Portuguese. FLiP includes a grammar checker, a spell checker, a thesaurus and a hyphenator that enable different proofing levels – word, sentence, paragraph and text – of European and Brazilian Portuguese. An online version is available at http://www.flip.pt/online.

[3] This is still an ongoing work to be detailed in future publications.

Since this clearly emulates an interactive task, our approach requires that the system processes one question at a time, which means that the topic detection is performed regardless of any information contained in the subsequent questions of the cluster. This requirement introduces a major problem to one-topic based approaches, i.e. those in which only one topic is considered, because it is often difficult to infer just from the first question which will be the topic of the whole set of questions. For example, consider the question 58 of the Spanish test set, "¿Quién diseñó el procesador Zilog Z80?" [Who designed the Zilog Z80 processor?], whose answer is "Federico Faggin". Only after parsing the subsequent question, "¿De cuántos bits era este procesador?" [How many bits was this processor?] could the system have inferred that the topic was "Zilog Z80", and not "Federico Faggin". To overcome this inherent difficulty, we adapted the system to be able to accept *several* topics, when more than one is possible. This was done by implementing the following procedure:

- The system parses the first question of a cluster and follows the usual procedure to extract an answer (if any is available);
- The system calls the method `GetTopic`, which collects the noun phrases (nouns, proper nouns and named entities) of the extracted answer and the noun phrases of the question which were considered object pivots (see section 2.1), merging these two groups into a single list of *topic pivots*;
- In subsequent questions of the same cluster, the system first parses the question, obtaining its pivots, and then calls the method `SetTopic`, which appends to the question pivots the list of topic pivots collected in the previous step.

We illustrate this procedure in Table 1 using the questions mentioned above.

Question	"¿Quién diseñó el procesador Zilog Z80?"
Pivots	*diseñar*, *procesador*, *zilog z80*
Extracted Answer	*Federico Faggin*
Topic pivots after `GetTopic`	*federico faggin*, *zilog z80*
Question	"¿De cuántos bits era este procesador?"
Pivots after `SetTopic`	*bits*, *procesador* + *federico faggin*, *zilog z80*
Extracted Answer	*8*

Table 1 – Procedure for topic-related questions.

Notice that our approach excludes anaphoric analysis of the questions. Naturally, the system will sometimes perform poorly as a consequence of an eventual excess of topic pivots. We are considering as future work a more sophisticated approach that involves anaphora resolution and through which we can do some sort of topic disambiguation by choosing only the co-references (i.e. the topics) that best suit the question. In the above example, the expression "este procesador" could only refer back to "Zilog Z80" and not to "Federico Faggin"; hence the topic pivot "Federico Faggin" could have been discarded if an anaphora resolution method had been used to disambiguate the topic.

2.3 Addition of the Wikipedia collection

Unlike the collections of newspaper articles, the Wikipedia collection has a rich structure (links, categories, disambiguation pages, etc.) that suggests using strategies capable of extracting knowledge from structured data. However, as a first approach, we did not use such sophisticated strategies; instead, we indexed the Wikipedia articles as natural language text and did some minor adaptations.

The indexation module ignores all the metadata included in tables and boxes. We did not index disambiguation pages, discussion pages, or any internal Wikipedia pages whose title starts with "Wikipedia:". We converted links of the form "[[<Article title>|<Link text>]]" into strings of the form "<Link text> (<Article title>)" (i.e., putting the title of the linked article between parentheses) and indexed it as natural text. This strategy allows answering some short definition questions (e.g. acronyms) by identifying the link text with the article title without making any change in the system modules.

We developed a simple scheme of anaphora resolution for Wikipedia articles. Since many anaphoric references in these articles have the article title as their referent, we adopted the following procedure: every time

we parse a sentence during the sentence retrieval stage and find a null subject or a personal pronoun subject, we replace it by the article title and parse the sentence again (of course, this approach excludes other possible referents besides that expressed in the article title, which is a limitation). Consider the question 136 of the Spanish test set "¿Cuánto mide de alto la Pirámide del Sol de Teotihuacan?" [How high is the Pyramid of the Sun at Teotihuacan?]. The following sentence appears in the article titled "Pirámide del Sol (Teotihuacan)" and the parser detects that it has a null subject: "Tiene 65 m de altura.". So a new sentence is composed, "Pirámide del Sol (Teotihuacan) tiene 65 m de altura", and the answer *65 m* is successfully extracted. Unfortunately, as we will see in section 3, some QA@CLEF evaluators considered these answers unsupported (even though the article title is provided) and this had a significant impact on the evaluation of our system in the Spanish run.

2.4 Changes in the processing modules and resources

Apart from adapting the system to the new requirements of the 2007 QA@CLEF edition, we did some internal changes in the processing modules, specifically with respect to the way QPs, APs and QAPs are parsed (see [2, 4] for more details). We have adapted Earley's parsing algorithm [5] to be able to handle our grammar for QA. This allowed us to introduce some new features that take profit of grammar recursion. For instance, a new `Rep` command was introduced to deal with the arbitrary repetition of a term (for instance, `Rep [Cat(N) Cat(Vg)]` stands for an arbitrary sequence of nouns followed by a comma). Amongst other things, this feature provides an efficient method for extracting answers from list questions. Another feature that uses recursion is the ability to follow different paths to ignore or to take into account text between parentheses, when testing a pattern. For example, consider the sentence "Jorge Sampaio (presidente de Portugal) deslocou-se em visita de estado à República Popular da China" [Jorge Sampaio (president of Portugal) has paid a state visit to the People's Republic of China]. The text between parentheses could be extracted as an answer to the question "Quem é Jorge Sampaio?" [Who is Jorge Sampaio?]. But if the question is "Que país visitou Jorge Sampaio?" [Which country did Jorge Sampaio visit?], it would be useful to ignore this portion of text. Using our adaptation of Earley's parser, both paths are explored when searching for an answer. This turns out to be particularly useful for Wikipedia articles, since we convert links to other articles into the article titles between parentheses (see section 2.3).

Finally, the Spanish language resources were enhanced with the improvement of the lexicon – by means of the introduction of many new words, inflections and derivations –, the recognition of named entities and the inclusion of a Spanish thesaurus.

3 Results

In the tables below, the sets of questions were classified according to three question categories: *factoid* (FACT), *definition* (DEF) and *list* (LIST), with the judgments used for evaluation (R=Right, W=Wrong, X=Inexact, U=Unsupported), as defined in the CLEF 2007 guidelines.

Whereas for Portuguese (PT) the results shown in Table 2 are those provided by the QA@CLEF evaluators[4], for Spanish (ES) we present instead our internal evaluation. This is justified by the fact that we strongly disagree with the Spanish QA@CLEF assessors that have considered 20 answers as unsupported (U) instead of right (R), having a negative impact of 10% in the accuracy of our system. Most of these answers are directly related to the procedure described in the previous section 2.3; in other languages (e.g. Portuguese) the assessors considered such answers as R in similar situations. It seems to us that such a discrepancy between our judgement and that of the Spanish assessors is relevant enough to be discussed in the present paper. We illustrate with two examples: *(i)* the question 136 referred in section 2.3, "¿Cuánto mide de alto la Pirámide del Sol de Teotihuacan?" [How high is the Pyramid of the Sun at Teotihuacan?], for which our system extracted the answer *65 m* from the Wikipedia article "Pirámide del Sol (Teotihuacan)" and provided the snippet "Tiene 65 m de altura", and *(ii)* the question 179 "¿Cuántos récords mundiales batió?" [How many world records did he achieve?], with the topic *Johnny Weissmüller*, for which our system answered *67* and provided the snippet "Ganó 52 campeonatos naciones [sic] de Estados Unidos y estableció un total de 67 récords mundiales", from the Wikipedia article "Johnny Weissmüller". Both answers were considered unsupported by the Spanish

[4] The results for Portuguese include 5 X answers that the Portuguese CLEF assessors marked with M, an unknown classification according to the CLEF guidelines, and the total of answers accounts for the 200 we provided instead of the 195 accounted in the CLEF organisation statistics.

assessors, but there seems to be no clear support of that in the QA@CLEF guidelines which are open to different interpretations, by stating that:

> "Each exact answer must be supported by:
> - the DOCID of the document in the news collection or by the filename of the dumped Wikipedia (November 2006) page, from which it has been retrieved;
> - portion(s) of text, which provide enough context to support the correctness of the exact answer. Supporting texts may be taken from different sections of the relevant documents, and which must sum up to a maximum of 700 bytes. Unnecessarily long snippets, i.e. those that do not meet this requirement, might be judged as non-supporting."

Under our criteria, the Spanish results considered in Table 2 are significantly higher than CLEF's official evaluation, which has determined an overall accuracy of 44.5%.

	Q \ A	R		W		X		U		Total		Accuracy	
		PT	*ES*	*PT*	*ES*	*PT*	*ES*	*PT*	*ES*	*PT*	*ES*	*PT*	*ES*
Non-topic-related	FACT	63	77	43	57	3	2	1	1	110	137	**57.3%**	**56.2%**
	DEF	25	19	2	7	5	0	0	0	32	26	**78.1%**	**73.1%**
	LIST	4	2	4	5	0	0	0	0	8	7	**50.0%**	**28.6%**
	Total	92	98	49	69	8	2	1	1	150	170	**61.3%**	**57.6%**
Topic-related	FACT	8	11	39	16	2	1	0	0	49	28	**16.3%**	**39.3%**
	DEF	0	0	0	0	0	0	0	0	0	0	-	-
	LIST	0	0	1	2	0	0	0	0	1	2	-	**0.0%**
	Total	8	11	40	18	2	1	0	0	50	30	**16.0%**	**36.7%**
General (All)	FACT	71	88	82	73	5	3	1	1	159	165	**44.7%**	**53.3%**
	DEF	25	19	2	7	5	0	0	0	32	26	**78.1%**	**73.1%**
	LIST	4	2	6	7	0	0	0	0	9	9	**44.4%**	**22.2%**
	Total	100	109	90	87	10	3	1	1	200	200	**50.0%**	**54.5%**

Table 2 – Results by category of question, including detailed results of topic and non topic-related questions.

The general results of Table 2 show that there is a significant decrease of overall accuracy in PT and an increase in ES when comparing with the performance of the system in the last CLEF campaign. In order to compare the current results with the results of previous editions of QA@CLEF, we analysed separately the accuracy of non-topic-related questions, taking into account just those questions that could be directly answered without the need to analyse relations between previous questions or answers.

An analysis of the question clusters in the Portuguese and Spanish sets brings some noticeable differences within their number and size. While Portuguese had 25 clusters including 75 questions, Spanish had 20 clusters including 50 questions. In addition to this, while the majority of Spanish clusters had two questions (13 two-question, 4 three-question and 3 four-question clusters), the Portuguese set had as many four-question as two-question clusters (10 two-question, 5 three-question and 10 four-question clusters). These figures show a higher complexity of the Portuguese set of questions, which could have led to its lower results. Beside this, the Portuguese set also presented a few elliptic questions, such as questions 33, "Onde?" [Where?], 68, "E Régulo?" [And Regulus?], 104, "Por quem?" [By whom?] and 184, "Quando?" [When?].

The inclusion of Wikipedia articles in this year's target collections increased the level of difficulty, since previously only newspaper corpora were used. The huge volume of information from the combined corpora (news collection and Wikipedia) was in some cases a handicap to reach the best answers. Syntactic processing helped us, though, to better tune the categorisation of the questions and to structure the information provided by the question pivots.

With regard to non-topic-related questions, the system responded quite satisfactory, both in PT and in ES runs. Besides that, general improvements in the Spanish modules raised the accuracy of the ES run. On the other hand, the accuracy of topic-related questions is by far lower in both languages. Once more, the above mentioned differences between both language question sets can be seen as a significant reason for the stronger impact on Portuguese results.

Table 3 displays the distribution of errors along the main stages of Priberam's QA system:

Stage ↓ Question →	W+X+U		Failure (%)	
	PT	ES	PT	ES
Document retrieval	45	16	45.0	17.6
Extraction of candidate answers	23	37	23.0	40.7
Choice of the final answer	21	29	21.0	31.9
NIL validation	4	6	4.0	6.6
Other	7	3	7.0	3.3
Total	100	91	100.0	100.0

Table 3 – Reasons for W, X and U answers

In the ES run, the reasons for wrong answers do not represent a big change in comparison with last year's percentages. As in CLEF 2006 results [4], most errors occur during the extraction of candidate answers. In the PT run, most errors occur during the document retrieval stage, especially in the topic-related questions (60% of the errors). In some cases, a possible explanation is that the topic may be erroneously detected, resulting in too many pivots being assigned to the subsequent questions, which causes the system to miss the documents in which the correct pivots appear.

Some of the wrong answers classified as *Other* are due to dubious clusters of questions that seem not to respect the guidelines, since topics do not always come from the first question/answer pair, as illustrated in Table 4. Notice the change of topic from "Bill Gates" (question ids 185 and 186 of the ES set) to "Universidad de Harvard" (question ids 187 and 188).

Id	Group id	Question
185	2161	¿Cómo se llama la mujer de Bill Gates? [What is the name of Bill Gates' wife?]
186	2161	¿En qué universidad estudiaba él cuando creó Microsoft? [At which university was he studying when he created Microsoft?]
187	2161	¿Qué presupuesto tenía esa universidad en 2005? [What was the budget for that university in 2005?]
188	2161	¿En que [sic] año se fundó? [In which year was it founded?]

Table 4 - Example of a double-topic cluster of questions.

There are some examples too, in both languages, of answers that apparently can only be extracted from tables or boxes, such as the coordinates of Guarda for the PT question 87 "Qual é a latitude e longitude da Guarda?" [What are the latitude and longitude of Guarda?] or the distribution company for the ES question 172 "¿Cuál es la distribuidora de la película El Planeta de los Simios que se estrenó en 1968?" [Who was the distributor for the 1968 movie Planet of the Apes?]. In some cases, the documents were retrieved, but the answer was not extracted.

To sum up, in the analysis of this year's results, we find it somewhat complicated to evaluate the introduction of two simultaneous major changes in the same CLEF campaign: is the problem in the way the system handles the test set or in the way the system reads the text collection? Gradual changes are easier to assess and are more helpful for the participants, allowing them to correct errors and try different approaches, hence contributing to the development of better and more robust systems. Besides, any comparative analysis of the system's performance in multiple languages is made impossible due to the multiple criteria of the different assessors while interpreting the guidelines.

4 Conclusions and future work

Due to the changes that occurred in this year's QA@CLEF task, we cannot directly compare the accuracy of the Priberam's QA system with previous CLEF campaigns. The system was adapted to handle the Wikipedia collection and to answer topic-related questions recurring to simple approaches given the short amount of time available to prepare our participation; we believe that more sophisticated strategies would lead to a much better performance, but would also require extra time.

We can say, however, that Priberam's system achieved a more accurate question categorisation, hence decreasing the number of wrong candidate answers, and this was due to the introduction of syntactical parsing during question processing. We expect to further improve the syntactical analysis and to extend it to the answer extraction module. The answer syntactic analysis will allow the system to more precisely match the pivots of the question with their counterparts in the answer, taking into account their syntactic functions. A further development of anaphora resolution will also be one of our goals in the future. We expect to broaden the approach applied this year in Wikipedia article titles, by using our syntactic parsing engine to deal with co-references also in text sentences.

Finally, we intend to evaluate again the cross-language performance of our system, as this was left out in this campaign. In particular, it would be interesting to evaluate how the system performs with topic-related questions in a cross-language environment.

Acknowledgments

Priberam Informática would like to thank Synapse Développement, TiP, University of Economics of Prague (UEP), as well as the CLEF organisation and Linguateca. We would also like to acknowledge the support of the European Commission in the M-CAST (EDC 22249 M-CAST) project.

References

[1] Vallin A., B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peas, M. de Rijke, B. Sacaleanu, D. Santos, R. Sutcliffe (2005), Overview of the CLEF 2005 Multilingual Question Answering Track, *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop* (CLEF 2005) (Vienna, Austria, 21-23 September).
Available at: http://clef.isti.cnr.it/2005/working_notes/workingnotes2005/vallin05.pdf

[2] Amaral C., H. Figueira, A. Martins, A. Mendes, P. Mendes, C. Pinto (2005), Priberam's question answering system for Portuguese, *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop* (CLEF 2005) (Vienna, Austria, 21-23 September).
Available at: http://www.clef-campaign.org/2005/working_notes/workingnotes2005/amaral05.pdf

[3] Magnini B., D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, A. Peñas, P. Rocha, B. Sacaleanu, R. Sutcliffe, Overview of the CLEF 2006 Multilingual Question Answering Track, *Working Notes for the CLEF 2006 Workshop* (CLEF 2006) (Alicante, Spain, 20-22 September).
Available at: http://www.clef-campaign.org/2006/working_notes/workingnotes2006/magniniOCLEF2006.pdf

[4] Cassan A., H. Figueira, A. Martins, A. Mendes, P. Mendes, C. Pinto, D. Vidal (2006), Priberam's question answering system in a cross-language environment, *Working Notes for the CLEF 2006 Workshop* (CLEF 2006) (Alicante, Spain, 20-22 September).
Available at: http://www.clef-campaign.org/2006/working_notes/workingnotes2006/cassanCLEF2006.pdf

[5] Earley, J. (1970), An Efficient Context-Free Parsing Algorithm. In Communications of the ACM, 13(2): pp. 94-102.