Priberam's question answering system in QA@CLEF 2008

Carlos Amaral, Adán Cassan, Helena Figueira, André Martins,
Afonso Mendes, Pedro Mendes, José Pina, Cláudia Pinto

Priberam
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, jfp, cp}@priberam.pt

Abstract

This paper describes the changes implemented in Priberam's question answering (QA) system since our last QA@CLEF participation, followed by the discussion of the results obtained in Portuguese and Spanish monolingual runs at the main task of QA@CLEF 2008. This time, the main goal of Priberam's participation, following the results of last year's evaluation, was to stabilize the system in order to achieve its potential performance. To attain that performance status, we enhanced the syntactic analysis of the question and improved the indexing process by using question categories at the sentence retrieval level and ontology domains of the expected answer in document retrieval. The fine-tuning of the syntactic analysis, by defining and using core nodes of phrases as objects, allowed the system to more precisely match the pivots of the question with their counterparts in the answer, taking into account their syntactic functions. As a result, in QA@CLEF 2008, Priberam's system achieved a considerable overall accuracy increase in the Portuguese run.

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.

Keywords
Question answering, Questions beyond factoids, Query Expansion, Portuguese, Spanish.

1 Introduction

The performance of Priberam's system in last year's QA@CLEF displayed internal and external changes. Internally, the system underwent several modifications, both in the Portuguese and in the Spanish modules, the most relevant one being the introduction of syntactic question processing [1]. Externally, the CLEF organisation introduced topic-related questions (questions clustered around a common topic that might present anaphoric links between them) and added Wikipedia as a target document collection to the already existent newspaper corpora [2]. As a result, there was a slight increase of the overall accuracy in the Spanish (ES) run and a significant decrease of the overall accuracy in the Portuguese (PT) run. Nevertheless, Priberam's system achieved a more accurate question categorisation, hence decreasing the number of wrong candidate answers, due to the introduction of syntactic parsing during question processing.

The main goal of Priberam's participation in QA@CLEF 2008 was to stabilize the system in order to surpass the results it obtained in previous QA@CLEF participations [3, 4]. To enhance its performance, we improved the indexing/retrieval process by using question categories (QC) at sentence retrieval level and

ontology domains of the expected answer in document retrieval. The fine-tuning of the syntactic analysis, by using the phrases' core nodes as objects, allowed the system to more precisely match the pivots of the question with their counterparts in the answer, taking into account their syntactic functions. As a result, in QA@CLEF 2008, Priberam's system achieved a considerable overall accuracy increase in the Portuguese run.

The paper is organised as follows: section 2 describes the major adjustments made to the system, such as the work done in improving the syntactic processing of the question and the adaptations to deal with topic-related questions; section 3 analyses and discusses the results of both monolingual runs; section 4 presents the conclusions and future work.

2 Adaptations and improvements of the system

Priberam's QA open-domain system has already been described in detail [3, 5]. Briefly, it relies on a set of linguistic resources (such as a wide coverage lexicon, a thesaurus and a multilingual ontology) and software tools (which can be used to write and test grammars, to build contextual rules for performing morphological disambiguation or named entity (NE) recognition, to build patterns for question categorization/answer extraction, etc.). This general domain QA system is based on 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 (earlier 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.

Since the overall architecture of Priberam's QA system remains unchanged, this year we focused on (i) the improvement of the indexing/retrieval process, (ii) the refinement of the question syntactic analysis, (iii) the fine-tuning of named entity recognition (NER), and we also revised (iv) the treatment of topic-related questions.

2.1 Improvements of the indexing/retrieval process

This year we kept the approach used and described on previous CLEF campaigns [3], but the system was submitted to a lot of fine-tuning and optimization in order to improve performance. Some of the enhancements allowed us to go further on what we indexed and queried for without major performance penalties. The most important changes were indexing of QCs at sentence level instead of at document level, the complete indexation of ontology domains at document level and the use of different ratings for document titles and document body (both for Wikipedia and newspaper articles).

In [3] we described the work done in two different steps, document retrieval and sentence retrieval. Much of the work done on the second step is now also done on the first step because many of the problems the system experienced in the retrieval process were due to the loss of documents in document retrieval.

The following summarizes the most important changes implemented:

1) It is now possible to embed in the QAPs rules for querying the ontology of the target answer (see section 2.2);
2) A document indexed with the QC on the same sentence as the pivots has now a much higher rating;
3) Documents where the pivots (especially NEs) appear in the title have priority over the other documents;
4) Documents that are more recent have higher priority (this is relevant for news corpora);
5) It is now possible to write rules to tag some pivots with higher/lower priority or discard them for retrieval.

Both in the Portuguese and Spanish runs the changes proved to be very rewarding, since the failures due to the retrieval stage dropped from PT-45%, ES-17.6% last year to PT-4.1%, ES-0.9%.

From the analysis of the four failures during document retrieval stage on the Portuguese run, we can see that two were due to bad handling of date restrictions. In the PT question 75 “Quantos tem hoje em dia?” [How many does it have nowadays?], whose topic is *Berlim,* the system did not translate the expression “hoje em dia” [nowadays] to the current date. In the PT question 6 “Diga uma escola de samba fundada nos anos 40.” [Say one samba school founded during the 40s] the date expression “anos 40” [the 40s], which should be selecting all documents with dates between 1940 and 1950, was wrongly only selecting documents with dates from 1940. The other two failures were due to the presence of very common pivots with very frequent QCs for those pivots. Even though this did not happen very often this year, it is probably the main cause of errors in the retrieval stage.

On the Spanish run there was only one failure during the retrieval stage. ES question 112 “En qué año la construyeron?” [In which year was it built?] failed because the sentence containing the answer “en lo alto de la fachada está grabada la fecha de su construcción: 1539” was not being indexed with QC <DATE>.

Even though the results were very good for the retrieval stage, some improvements still need to be addressed in the future: (i) a new schema for indexing and querying date periods and (ii) a new schema for indexing QCs, where we plan to tag each word or phrase in the indexing process with QCs, instead of indexing QCs by sentence. For instance, in the sentence “Manuel II de Portugal, último rei de Portugal de 1908 a 1910” we should index the noun phrases “Manuel II de Portugal” and “último rei de Portugal” with QCs <CHRONOLOGY> and <FUNCTION> and “último rei de Portugal” with QCs <CHRONOLOGY> and <DENOMINATION>.

2.2 Refinement of the question syntactic analysis

During the question analysis stage, questions are categorised and syntactically parsed. We maintain the approach presented last year, which introduced the possibility to capture the syntactic structure of the question by using FLiP's linguistic technology[1] [1]. The main difference is that now we detect the core nodes of the syntactic phrases and use them as the question's objects (its main constituents).

Each syntactic phrase may have one or more core nodes, that may coincide with the head phrase or not, and that are assigned to different object types accordingly to their relevance in extracting the expected answer. Object assignment is done after parsing, using the syntactic information that was treated in that stage. Typically, object assignment establishes a hierarchy of objects: it places the core nodes of subjects at the top, followed by those of the verb's complements, the head of the verb phrase and the adjuncts. It also gives priority to NEs: for example, PT question 33 “Que político é conhecido como Iznogoud?” [Which politician is known as Iznogoud?] retains “Iznogoud” as the object, “é conhecido” as the verbal object and “político” as the restraining object.

This strategy can help solving a few simple instances of syntactic ambiguity, such as those derived from prepositional phrase (PP) attachment, in case of overgeneration or parsing errors [6], since the core nodes remain the same. For instance, in PT question 62 “Qual a largura do Canal da Mancha no seu ponto mais estreito?” [What is the width of the English Channel at its narrowest point?], which has three contracted prepositions (“do”, “da” and “no”), the parser could wrongly build the PP “do Canal da Mancha no seu ponto mais estreito” [of the English Channel at its narrowest point]. If the parser could not find its core nodes, the whole PP would be used as the object, thus introducing noise in the document retrieval stage. By establishing core nodes, one can assign the detected NE “Canal da Mancha” as the object and “no seu ponto mais estreito” as the modifying object.

We added a specific object, the interrogative object, which works as a placeholder for the expected answer. We use it along with the QC to narrow the search for target sentences and extract the answer. The use of its ontological domains led to a considerable increase in the accuracy of the retrieval process. For instance, in PT question 1 “Que tipo de animal é o Cocas?” [What kind of animal is Kermit?], the system looks for documents containing words and expressions belonging to the same ontology level of “animal”, the question's interrogative object. Thus, sentences that do not contain the word "animal", but contain words like “sapo” [toad] or “rã” (frog), are retrieved.

2.3 Fine-tuning of named entity recognition (NER)

The NER engine Priberam has been using in its QA system participated this year in the second edition of HAREM, an evaluation contest for Portuguese NER[2]. This participation led to an external evaluation of the engine and, consequently to its improvement and refinement. This had an impact on the precision of the answer extraction, namely in the more specific QCs. Besides the NEs already detected (e.g. people, places and organisations), we had to build new rules to recognize NEs that denote written and not written works, things (objects, substances), events, abstractions and numeric values (currencies, quantities, classifications). The rules that recognise time expressions were also improved, because Priberam’s NER engine was a participant in the time track of the second HAREM as well.

This, as mentioned above, was particularly important for some QCs such as <WRITTEN WORK>, <NOT WRITTEN WORK>, <STAR> or <CLASSIFICATION>. For QCs such as <DENOMINATION>, <FUNCTION> or

[1] FLiP, or *Ferramentas para a Língua Portuguesa*, is Priberam’s proofing tools package for Portuguese. FLiP includes a grammar and style 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.

[2] HAREM is organised by Linguateca; more information at http://www.linguateca.pt/HAREM/.

<LOCATION>, the semantic values of NEs were already being used in the indexing process and answer extraction, allowing the system to perform more accurately in these categories. With the addition of the new semantic tags and the creation of new rules that classify NEs using those tags, we were able to narrow the number of candidate answers in the more specific QCs. Thus, for a question such as topic-related PT question 162 "Diga um desses filmes." [Name one of those films.], whose topic is *Jean Vigo*, candidate answers that contained NEs classified as not written works were given a higher score.

Not only does this fine-tuning of the NER improve the answer extraction process, it also improves the syntactic parsing by restricting, for example, the number of PPs, hence preventing overgeneration, which will in turn create a more precise parser (see section 2.2).

The performance of Priberam's NER engine led to its commercial exploration: it is now being used for search refining in the sites of two major news media, *TSF* radio station[3] and *Jornal de Notícias* newspaper[4].

2.4 Dealing with topic-related questions

As mentioned on last year's working notes, the procedure for dealing with topic-related questions could perform poorly because of the excess of pivots. Moreover, since we just merged the question pivots, we loosed the question syntactical analysis. Like last year we only analyse the first question from the set and the current question, which means that we do not keep track of the changes to the topic. This had an impact on the Spanish questions but not on the Portuguese ones. In our opinion, topic-related questions are not very interesting for a commercial system at this stage of QA systems. Having this in mind, we developed the module only for CLEF and did not invest a lot of effort here.

The strategy we applied this year to topic-related questions was the following (see Table 1 examples):

1) analyse the first question;
2) save the answer;
3) analyse the current question;
4) handle explicit anaphors (those where the pronoun is expressed);
5) use the last expressed QC;
6) use the argument analysis of the question which expresses the QC;
7) import the missing arguments from the first question to the current question;
8) if the QC changes, also import the answer.

Question	QC	Objects	Answer
PT 11: Qual é a montanha mais alta do México? [Which is the highest mountain in Mexico?]	<MOUNTAIN>	- México - mais alta	Citlaltépetl
PT 12: E do Japão? [And in Japan?]	NIL	- Japão - NIL	
PT 12 final question analysis:	<MOUNTAIN> (inherited from PT 11)	- Japão (since it is expressed) - mais alta (inherited from PT 11)	(the system does not import the answer to PT 11 as an object because it has the same QC)
PT 81: Quem foi o último rei de Portugal? [Who was the last king of Portugal?]	<FUNCTION>	- último rei de Portugal	D. Manuel II
PT 82: Em que período foi ele rei? [In which period was he a king?]	<CHRONOLOGY>	- ele rei	
PT 82 final question analysis:	<CHRONOLOGY>	- D. Manuel II rei de Portugal	

Table 1 – Examples of question analysis of topic-related questions.

This procedure still has many flaws and systematically failed in questions like PT questions 37 "E um não-metal." [And a nonmetal.], 65 "E do pão?" [And of bread?] and 144 "E a segunda?" [And the second one?],

[3] http://www.tsf.pt.
[4] http://www.jn.pt.

where the arguments of the first question were not replaced but added. In the Spanish run, topic-related questions suffered with this new schema, since question syntactical analysis is still quite poor when compared to Portuguese.

3 Results

Table 2 presents the results of Portuguese and Spanish monolingual runs submitted by Priberam to the main task of QA@CLEF 2008. 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 CLEF 2008 guidelines.

	Q \ A	R		W		X		U		Total		Accuracy	
		PT	*ES*	*PT*	*ES*	*PT*	*ES*	*PT*	*ES*	*PT*	*ES*	*PT*	*ES*
Non-topic-related	FACT	83	55	24	45	4	0	1	2	112	102	**74.1%**	**53.9%**
	DEF	18	15	4	3	6	0	0	0	28	18	**64.3%**	**83.3%**
	LIST	3	5	3	8	3	4	0	1	9	18	**33.3%**	**27.8%**
	Total	104	75	31	56	13	4	1	3	149	138	**69.8%**	**54.3%**
Topic-related	FACT	23	11	23	46	1	1	3	1	50	59	**46.0%**	**18.6%**
	DEF	0	0	0	1	0	0	0	0	0	1	-	**0%**
	LIST	0	0	1	2	0	0	0	0	1	2	**0%**	**0%**
	Total	23	11	24	49	1	1	3	1	51	62	**45.1%**	**17.7%**
General (All)	FACT	106	66	47	91	5	1	4	3	162	161	**65.4%**	**41.0%**
	DEF	18	15	4	4	6	0	0	0	28	19	**64.3%**	**78.9%**
	LIST	3	5	4	10	3	0	0	1	10	20	**30.0%**	**25.0%**
	Total	127	86	55	105	14	4	4	4	200	200	**63.5%**	**43.0%**

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

Regarding the Portuguese run, the improvement of more than 20% in the accuracy of general factoid questions considerably contributed to the increase of the overall accuracy, which surpassed that of last year (50%). Besides that, an analysis of PT question clusters shows that there was an increase in the number of clusters (37) in a total of 88 questions, 51 topic-related, but that the system was able to extract the correct answers 45% of the times, which means a boost of nearly 30%, when comparing to last year's results.

Despite these general positive results, Table 2 also shows a decrease of accuracy in DEF and LIST questions. The reasons for failures are assembled in Table 3, which displays the distribution of errors, in both monolingual runs, along the main stages of Priberam's QA system. In the Portuguese run, the main source of error was the extraction of candidate answers, followed by the choice of the final answer. The main reason for errors in extraction of candidate answers is the coverage of QAPs, which are handwritten and therefore limited.

Stage ↓ Question →	W+X+U		Failure (%)	
	PT	ES	PT	ES
Document retrieval	4	1	4.1	0.9
Extraction of candidate answers	33	75	46.6	66.4
Choice of the final answer	20	17	27.3	15.0
NIL validation	8	9	11.0	8.0
Topic	4	7	5.5	6.2
Other	4	4	5.5	3.5
Total	73	113	100.0	100.0

Table 3 – Reasons for W, X and U answers

In the case of DEF questions, two of the extracted answers considered inexact by the CLEF assessors presented brackets (PT questions 27 and 88), here accounted as a problem of choice of the final answer. We opted to keep the text in brackets (the way the links in the Wikipedia articles were stored) because they normally present useful information to the user. At least one of the answers classified as *Other* in the PT run could be considered a possible assessor's error: in DEF question 137, whose topic is *Prémio Cervantes*, "Quem é que ganhou o prémio em 1994?" [Who won the prize in 1994?], the QA system extracted the answer "Mario Vargas Llosa" from the snippet "O escritor peruano Mario Vargas Llosa ganha o Prémio Cervantes 1994, o maior galardão literário espanhol." [Peruvian writer Mario Vargas Llosa wins Cervantes Prize 1994, the biggest Spanish literary prize]. Finally, the system did not extract any answer to DEF question 66 "O que é o jagertee?" [What is the jagertee?] because the snippet containing the correct answer was not indexed for unknown reasons and thus could not be retrieved.

With regard to the Spanish run, Table 2 shows that results within non-topic-related questions are quite similar to those of last year, while topic-related questions had a decrease in its accuracy of almost 20%. At this point, it deserves to be said that the number of both question clusters and topic-related questions doubled in 2008 for the ES test set: from 20 clusters and 30 topic-related questions, it passed to 48 clusters and 62 topic-related questions. This fact had, consequently, a strong impact on the Spanish results, both on the falling of non-topic questions accuracy by itself and, mainly, on the global results. Another remarkable fact about the Spanish set is a significant increase of the number of LIST questions compared to last year's set or to the Portuguese set.

In Table 3 we classified as *Other* all the unsupported answers in the ES run. All of them are certainly correct answers, but at least three of them do not explicitly contain all the needed supporting information in the snippet, although this information does appear in the document. Those errors could be seen as presentation errors, as a limitation of the system in the way of presenting the information, and not in the way it processes those questions. One of these examples is ES question 10 "¿A qué edad murió Wallace Rowling?" [At which age did Wallace Rowling die?]. The QA system correctly answered "67 años" from the snippet "- Sir Wallace Rowling, ex primer ministro de Nueva Zelanda, 67 años.". Although the actual snippet does not support the answer, the document where it comes from is a list of deceased people in 1995 from EFE, but that is not shown in the snippet. Something similar happens with ES question 170 "Según el plan Belloch, ¿cuántos vehículos policiales habrá en Barcelona?" [Acording to the Belloch plan, how many police vehicles will there be in Barcelona?] where the given answer is "301" from the snippet "Barcelona contará a partir del día 1 de enero con 301 vehículos policiales patrullando sus diez distritos: 114 por la mañana, 114 por las tardes y 73 por la noche." Again, the document does talk about the matter, the "Plan Belloch", even if the snippet does not.

Finally, there is an interesting case of extraction problem with the answer to the ES question 75 "¿Cómo se pronuncia eso?" [How is it pronounced?], whose topic is *TeX*. The correct answer is displayed in the Wikipedia between square brackets, and it happens to be ignored by the QA system because of that.

From the analysis of the results, we conclude that the retrieval stage and the question analysis stage are performing very well for questions like those posed in CLEF, that QAPs need to broaden their coverage and that the work done for Portuguese this year must be ported to the Spanish rules.

4 Conclusions and future work

Priberam's aim for QA@CLEF 2008 was to consolidate the system and improve its performance. Even though this year there was no real time exercise, from our tests we verified that we doubled the speed of the system and

improved the capacity to answer multiple questions simultaneously by enhancing the parallelism of the algorithms. These improvements were crucial for the implementation of the search engine in the sites of *TSF* and *Jornal de Notícias*. The retrieval stage is performing very well and the changes in the syntactical/semantic analysis now cover all the QCs.

During last year, we have been working on anaphora resolution and we had a first prototype of the system a few days before CLEF. As we had no feedback on how the system was behaving, we decided not to submit the results with anaphora resolution. After CLEF, we managed to run the tests and found out that the results were almost the same. This is an interesting result and we are convinced that this is due to the CLEF set of questions being extracted directly from what is written in the documents. Anaphora resolution is important only when dealing with Wikipedia articles between the body of the text and the title (this was already implemented last year).

The work done this year in the Portuguese module must be done in the Spanish module, specifically on the syntactical/semantic analysis of the questions and NER. As we mentioned in section 2.1, we plan to tag each word/phrase with the QC in the indexing stage. We hope that without a big penalty on index size we can achieve better accuracy and speed. Future work will also include, since we now have a big corpus of questions/answers/false answers, working on algorithms to automatically learn new question patterns from corpora.

Acknowledgments

Priberam would like to thank Synapse Développement, as well as the CLEF organisation and Linguateca.

References

[1] Amaral C., A. Cassan, H. Figueira, A. Mendes, P. Mendes, C. Pinto, D. Vidal (2007), Priberam's question answering system in QA@CLEF 2007, in A. Nardi and C. Peters (eds.), *Cross Language Evaluation Forum:Working Notes for the CLEF 2007 Workshop* (CLEF 2007) (Budapest, Hungary, 19-21 September).
Available at: http://www.clef-campaign.org/2007/working_notes/AmaralCLEF2007.pdf

[2] Giampiccolo D., A. Peñas, C. Ayache, D. Cristea, P. Forner, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu, R. Stutcliffe (2007), Overview of the CLEF 2007 Multilingual Question Answering Track, in A. Nardi and C. Peters (eds.), *Cross Language Evaluation Forum: Working Notes for the CLEF 2007 Workshop* (CLEF 2007) (Budapest, Hungary, 19-21 September).
Available at: http://www.clef-campaign.org/2007/working_notes/giampiccoloCLEF2007_Overview.pdf

[3] 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

[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, in A. Nardi, C. Peters, and J. Vicedo (eds.), *Cross Language Evaluation Forum: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] Amaral C., D. Laurent, A. Martins, A. Mendes, C. Pinto (2004), Design and Implementation of a Semantic Search Engine for Portuguese, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* (*Lisbon, Portugal, 26-28 May)*, vol. 1, pp. 247-250.
Also available at: http://www.priberam.pt/docs/LREC2004.pdf

[6] Zhao, S., D. Lin, (2005), A Nearest-Neighbor Method for Resolving PP-Attachment Ambiguity, in Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee, OiYee Kwong (eds.), *Natural Language Processing – IJCNLP 2004*, *First International Joint Conference, Hainan Island, China, March 22-24, 2004*, *Revised Selected Papers*, Berlin: Springer-Verlag, (LNCS n.º 3248), pp.545-554.
Also available at: http://www.cs.rochester.edu/~zhao/IJCNLP04-ppa.pdf