What Happened to Esfinge in 2007?

Luís Miguel Cabral, Luís Fernando Costa, and Diana Santos

Linguateca, Oslo node, SINTEF ICT, Norway
{luis.m.cabral,luis.costa,Diana.Santos}@sintef.no

Abstract. Esfinge is a general domain Portuguese question answering system which uses the information available on the Web as an additional resource when searching for answers. Other external resources and tools used are a broad coverage parser, a morphological analyser, a named entity recognizer and a Web-based database of word co-occurrences.

In this fourth participation in CLEF, in addition to the new challenges posed by the organization (topics and anaphors in questions and the use of Wikipedia to search and support answers), we experimented with a multiple question and multiple answer approach in QA.

Keywords: Question answering, Portuguese, anaphor resolution, question reformulation, answer choice, Wikipedia processing.

1 Introduction

This year's evaluation contest required the systems to adapt to two brand-new conditions: The difficulty of questions was raised by the introduction of topics and anaphoric reference between questions on the same topic; and the difficulty of answers was raised because collections included Wikipedia, in addition to the old newspaper collections. Our main goal this year was therefore to adapt Esfinge to work in these new conditions, which basically consisted in creating an initial module for creating non-anaphoric questions (resolving co-reference) to be input to (the previous year's) Esfinge, and a final module that dealt with the choice of multiple answers from several different collections and/or Esfinge invocations (multi-stream QA). As will be explained below, unexpected problems led us to also try a radically different approach based on a set of patterns obtained from the initial module.

2 Esfinge in 2007

Esfinge participated at CLEF in 2004, 2005 and 2006, as described in detail in the corresponding proceedings. Most work in Esfinge this year was related to address the new challenges introduces in QA@CLEF. Figure 1 gives a general overview of the system used this year.

There is a new `Anaphor Resolution` module to resolve anaphors, which adds, to the original question, a list of alternative questions where the anaphors are (hopefully) resolved. In addition, it may also propose relatively trivial reformulations. Then, for each of the alternative questions:

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 261–268, 2008.

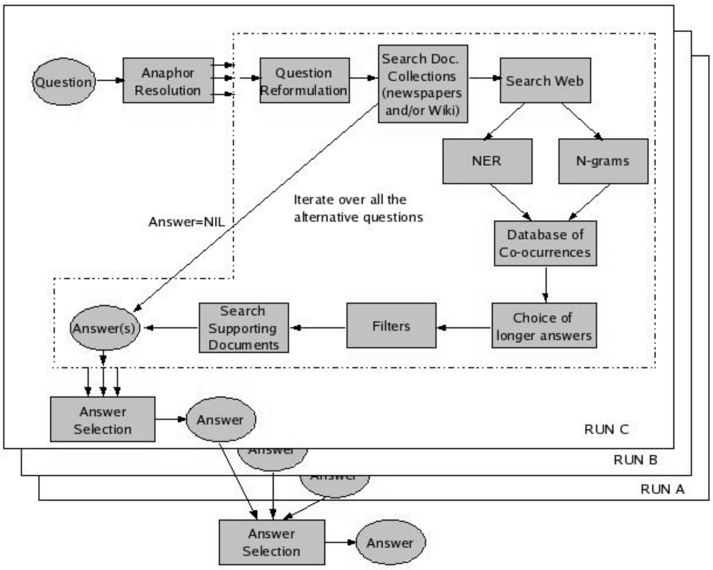


Fig. 1. Architecture of Esfinge 2007

1. The `Question Reformulation` module transforms the question into patterns of plausible answers. These patterns are then searched in the document collection using the `Search Document Collections` module. This module was adapted to allow search also in Wikipedia.
2. If the patterns are not found in the document collections, the system returns NIL and stops. Optionally, it can proceed by searching the same patterns in the Web. Then, all texts retrieved are analysed using a named entity recognizer (NER) system and an n-grams module in order to obtain candidate answers. The candidate answers are then ranked according to their frequency, length and the score of the passage from where they were retrieved. This ranking is in turn adjusted using the BACO database of co-occurrences [1] and the candidate answers (by ranking order) are analysed in order to check whether they pass a set of filters and to find a document in the collections which supports them.
3. From the moment Esfinge finds a possible answer for the question, it checks only candidate answers that include one of the previously found answers. It will replace the original answer if the new one includes the original answer, passes the filters and has documents in the collection that support it.

After iterating over all alternative questions, Esfinge has a set of possible answers. That is when the new module `Answer Selection` comes to play. This module attempts to select the best answer to the given question, which will be the final answer returned.

3 Anaphor Resolution

We developed a module relying crucially on the PALAVRAS parser [2] to replace anaphoric expressions into fully descriptive expressions (i.e., independently understandable questions).

This question reformulation is our first instantiation of the M,N-O,P model introduced in [3]. Basically, from the input question, we produce an (ordered) set of questions to be input to the original (one question, one answer) Esfinge system. Although this model does not cover everything required by interactive question answering, especially when user follow-up questions relate to previous answers and not to previous questions [4], question reformulation and choice among many answers was high on our research agenda.

The linguistic description of the kinds of anaphors catered for by our system can be found in [5], where we analyse the four sets of 200 questions which had Portuguese as source language. Incidentally, there were quite different kinds of questions depending on the (original source) language, suggesting that more attention should be paid to language differences [6].

In short, we deal rather successfully with (i) pronominal anaphor (subject, direct object, indirect object and oblique); (ii) possessive anaphor; (iii) demonstrative anaphor; and (iv) null subject anaphor; but completely missed, or had bad results, for (v) definite description anaphor; (vi) implicit anaphor; (vii) short questions (incidentally, only in the Portuguese monolingual set); and (viii) anaphoric reference to previous answers. Numbers and examples are in table 1.

PALAVRAS is a broad-coverage dependency parser for Portuguese which is used extensively by Linguateca projects since 1999, resulting in a set of programs to deal with its output described at the AC/DC project website.[1]

For anaphor resolution, our hypothesis was that most anaphoric related antecedents would be major constituents. (In fact, this was not confirmed by the data.) So, we set out using PALAVRAS for obtaining argument phrases.

By considering the particular question set, however, it soon became apparent that syntax alone was often not enough to assign the right argument structure. (See again [5] for details.) The simpler the questions, the less syntax is going to help. We have therefore used a set of heuristics – both prior to and after invoking PALAVRAS – to provide for more than one question formulation, to cope with these possible shortcomings.

For each question submitted to PALAVRAS, we get: (i) the anaphoric element and the phrase it is included in, and (ii) a list of possible candidates: all arguments mentioned within the same topic that include a proper name, all adjuncts with the same property, and all proper names and dates as well.

Anaphor resolution proper then proceeds by creating a set of new questions replacing the anaphor with all possible referent candidates. Often, no syntactic clue can help choose which candidate is most appropriate, as in *Quais eram os primeiros nomes dos dois irmãos Piccard? Qual deles ...* or *Qual o período*

[1] http://www.linguateca.pt/ACDC/

Table 1. Distribution of the several kinds of anaphors in the material: in parentheses is the subset which depends on the previous answer(s)

Kind	Example question	PT-PT	PT-DE	PT-ES	PT-FR	Total
subject pronoun	Quem é o dono delas? Quem era ele ?	14	19 (1)	6 (1)	19 (1)	58
personal pronoun	Quem é que o afundou em 1985?	1	1	1	0	3
demonstrative pronoun	Que (...) ao EEE quando este entrou em vigor? Quem é que dirige essa agência?	2 (1)	0	8 (1)	13 (1)	23
possessive pronoun	Qual era o seu verdadeiro nome?	7 (1)	3	4	6	20
null subject	Quantos habitantes tinha?	11 (1)	1	6	3	21
definite desc.	Quantos lugares tem o estádio?	4	7	3	5 (2)	19
implicit	Quem é o actor principal?	6	0	1	2 (1)	9
short questions	Onde?	6	0	0	0	6
other	cada, null object	1	0	1	0	2
Total		52	31	30	48	161

de gestação do ocapi? Qual o seu peso? where one would have to list all three possible noun phrases to get one reformulation right.

We assessed the performance of the anaphoric resolution module in detail, this time, differently from [5], considering also the cases which we had not considered during development. Table 2[2] provides the system evaluation, as opposed to algorithm evaluation, see [7] for the distinction. It is interesting to note that the Portuguese-only material was the hardest by far.

Table 2. Anaphor resolution performance for the 161 cases (158 questions)

	Number of questions	Correctly detected	Spurious	Undetected	Correctly resolved	Accuracy (resolved/all)
PT-PT	52 (51)	34 (33)	1	18	26 (25)	26/52 (50%)
PT-ES	30	24	2	7	17	17/30 (57%)
PT-DE	31 (29)	22 (21)	5	9	21	21/31 (68%)
PT-FR	48	38	2	10	31	31/48 (64%)
Total	161 (158)	118 (116)	10	44	95	95/161 (59%)

A by-product of the `AnaphorResolution` module was the identification, for each question, of the main verb, its arguments and its adjuncts, together with the possible entities for cross-reference coming from previous analyses inside the same topic. During the submission process, we decided to experiment also with this set of patterns (obtained from syntactic analysis) as an alternative to the original Esfinge patterns. These are called "PALAVRAS patterns" in the present paper. However, since no ranking algorithm was associated to them, their use has to be investigated further to discover how to employ them more judiciously.

[2] Three questions in the material had two different anaphors.

4 Searching Wikipedia

The use of Wikipedia presented a new challenge for Esfinge. Fearing that the size of the text would make the current methods prohibitively slow (the initial downloaded size amounted to about 5.4G), we chose to store the text in a MySQL database, instead of compiling the text in the IMS-CWB. The process was similar to the one used in BACO, using indexing capabilities to allow faster queries on the collection, indexing words up to a minimum length of 3 characters, and storing the text in sets of several sentences instead of storing the entire article together. In order to keep the sentences' context, information was repeated, intercalating the sentences, instead of simply grouping consecutive sentences, as shown in table 4 of [5].

Having completed the preparation of the data for analysis, the next step consisted in making this data accessible to Esfinge, which was easy, given that Esfinge already used BACO's interface to MySQL that assessed rarity of words, as detailed in [8].

The main task was to make the Wikipedia collection work as just one more resource from which answers could be retrieved, independently of the implementation.

Esfinge generates several text patterns from the given question. Each one is then used to search within the collections. While Esfinge, previously, only catered for CQP patterns to be directly applied to the newspaper collections, corresponding patterns for the MySQL function `Match Against` had to be created to access the indexed text of Wikipedia.

While in CQP Esfinge produces several queries from one expression and later joins the results, in MySQL this was transformed into one single query, independent of word order. For example, the expression *+navegação +cabotagem* matches against the following sentence: *A* ***cabotagem*** *se contrapõe à* ***navegação*** *de longo curso....*

5 Choosing among Several Answers

For each question reformulation we had one answer, therefore the `Answer selection` module had to choose the final one. Also, we created a large number of runs with different options, employing different search patterns and using different textual resources. Initially we had run the following runs:

- One run with all collections (Web+News+Wiki),
- One run without consulting the Web (News+Wiki),
- One run without the Wikipedia collection (Web+News).

Later we ran the same options but used instead the patterns generated using PALAVRAS.

As we had only two possible runs to send, we used this module also to merge the results of the individual runs. We merged all runs that used the same kind of search patterns (Esfinge or PALAVRAS).

Merging took into consideration the sum of the following aspects: (i) the number of times a certain answer was found in all runs; and (ii) the relevance of the support text to the question asked, computed as the number of times that words (with 3 or more characters) in the question occurred in the support text.

To evaluate this module, we looked into the 378 cases (distributed over the 3 automatic selection runs, presented in Figure 2, as no. 6, 11 and 13) where the choice module had more than one non-NIL answer to choose from, and counted the cases where the right answer was among the candidates (80). For this number, the choice was right in 68.75% of the cases.

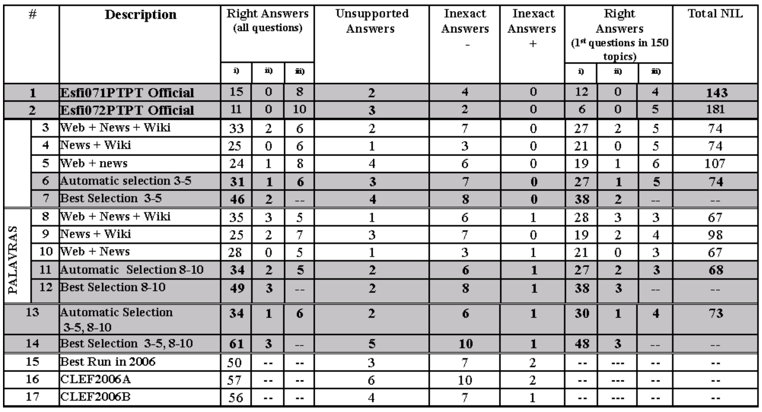
6 Our Participation and Additional Experiments

The official results can be seen in the first two lines of Figure 2, together with their subsequent repetition, after several severe bugs were discovered – unfortunately too late to resubmit to CLEF. Figure 2 displays the results of the individual runs and of their combination.

In order to assess the import of the `Answer Selection` module, we did a manual choice run as well (choosing manually among the different answers). This is indicated as **best** selection vs. **automatic** selection.

In order to evaluate the impact of adding Wikipedia as an additional source of knowledge, we also ran last year's questions with the new architecture (2006A and 2006B, respectively with Esfinge or PALAVRAS patterns), which resulted in a 3-4% improvement only.

Table 3 summarizes the main causes for errors in the best individual run (no. 8). The two main causes for wrong answers are both related to the retrieval of relevant documents. The category "Wrong or incomplete search patterns" refers to questions where the search patterns did not include the necessary information to answer the questions, while "Document retrieval failure" counts the cases



	#	Description	Right Answers (all questions)			Unsupported Answers	Inexact Answers -	Inexact Answers +	Right Answers (1st questions in 150 topics)			Total NIL
			i)	ii)	iii)				i)	ii)	iii)	
	1	**Esfi071PTPT Official**	15	0	8	**2**	4	0	12	0	4	**143**
	2	**Esfi072PTPT Official**	11	0	10	**3**	2	0	6	0	5	181
	3	Web + News + Wiki	33	2	6	2	7	0	27	2	5	74
	4	News + Wiki	25	0	6	1	3	0	21	0	5	74
	5	Web + news	24	1	8	4	6	0	19	1	6	107
	6	Automatic selection 3-5	**31**	**1**	**6**	**3**	7	**0**	27	**1**	**5**	**74**
	7	Best Selection 3-5	**46**	**2**	--	**4**	**8**	**0**	**38**	**2**	--	--
PALAVRAS	8	Web + News + Wiki	35	3	5	1	6	1	28	3	3	67
	9	News + Wiki	25	2	7	3	7	0	19	2	4	98
	10	Web + News	28	0	5	1	3	1	21	0	3	67
	11	Automatic Selection 8-10	**34**	**2**	**5**	**2**	**6**	**1**	**27**	**2**	**3**	**68**
	12	Best Selection 8-10	**49**	**3**	--	**2**	**8**	**1**	**38**	**3**	--	--
	13	Automatic Selection 3-5, 8-10	**34**	**1**	**6**	**2**	**6**	**1**	**30**	**1**	**4**	**73**
	14	Best Selection 3-5, 8-10	**61**	**3**	--	**5**	**10**	**1**	**48**	**3**	--	--
	15	Best Run in 2006	50	--	--	3	7	2	--	---	--	--
	16	CLEF2006A	57	--	--	6	10	2	--	---	--	--
	17	CLEF2006B	56	--	--	4	7	1	--	---	--	--

Fig. 2. Results of the additional experiments (A: Right answers including NIL; B: Partially right answers on lists; C: Right NIL answers)

where no relevant documents were retrieved in the collections, even though the search patterns included the necessary information. "Other" covers all causes that occurred less than five times.

In this table we counted the first module to fail. This explains why the initial modules are the ones with more errors: the modules which appear later are not even invoked for a significant part of the questions. Even though incompleteness of the search patterns was the main single cause for failure, this was to some extent due to poor communication among some modules (that was discovered only afterwards). It is important to point out that, still, the best run obtained by Esfinge used the PALAVRAS patterns.

7 Discussion and Further Work

We believe that the comparison of Esfinge results in 2006 and 2007 lends support to the claim that this year the difficulty of questions was raised, and we welcome this. Having the questions grouped in topics and including several types of anaphors brings us a step closer to the way humans ask questions and allowed us to develop Esfinge towards higher usefulness.

However, we think that the question set had too many errors to be used as a fair evaluation resource, and we hope that this won't be repeated in future editions of QA@CLEF.

This year, we concentrated mainly on developing the anaphor resolution module and the module responsible for merging and/or choosing from several alternative answers.

There is a lot of improvement that we can foresee for the first module, although a specific analysis of what errors are due to PALAVRAS performance as opposed to anaphor resolution proper is still due.

The choice algorithm also deserves closer attention, since it attained only 67% or 69% of the best combinaton when merging 3 runs, and 55% when it tried to merge all runs, producing in fact worse results than some of the individual runs it combined.

To deal with this, we are currently investigating several strategies: (i) to give different weights to different sources, (ii) combine the individual weights that

Table 3. Causes for wrong answers in the best individual run

Wrong	Answers
Co-reference resolution	25
Wrong or incomplete search patterns	63
Document retrieval failure	33
Mistake of the answer scoring algorithm	24
Mistake in the supported answer filter	7
Other	13
Total	165

had been assigned in each individual run, and/or (iii) saving more information, such as the patterns used to find each answer, for aiding the decision.

Acknowledgements. This work was done in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC.

References

1. Sarmento, L.: BACO - A large database of text and co-occurrences. In: Calzolari, N., et al. (eds.) Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 22-28 May 2006, pp. 1787–1790 (2006)
2. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)
3. Santos, D., Costa, L.: QolA: fostering collaboration within QA. In: Peters, C., et al. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 569–578. Springer, Heidelberg (2007)
4. Bertomeu, N., Uszkoreit, H., Frank, A., Krieger, H.U., Jörg, B.: Contextual phenomena and thematic relations in database QA dialogues: results from a Wizard-of-Oz Experiment. In: Proceedings of the HLT-NAACL 2006 Workshop on Interactive Question Answering (2006)
5. Cabral, L.M., Costa, L.F., Santos, D.: Esfinge at CLEF 2007: First steps in a multiple question and multiple answer approach. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop, Budapest, Hungary, 19-21 September (2007)
6. Santos, D., Cardoso, N.: Portuguese at CLEF 2005: Reflections and Challenges. In: Peters, C., ed.: Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop, Vienna, Austria, 21-23 September (2005)
7. Mitkov, R.: Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In: Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2000), Lancaster, UK, pp. 96–107 (2000)
8. Costa, L.: Question answering beyond CLEF document collections. In: Peters, C., et al. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 405–414. Springer, Heidelberg (2007)