Esfinge at CLEF 2007: First steps in a multiple question and multiple answer approach

Luís Miguel Cabral, Luís Fernando Costa, 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 analyzer, 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. Although the official results were severely compromised by a series of bugs, later experiments showed that the hardest – and so far mostly unsuccessful – subtask for Esfinge with several competing answers was to effect a principled choice among them. Anyway, access to Wikipedia managed to achieve better results on last year's questions, and, based on a satisfactory evaluation of the anaphoric reference module, we can conclude that Esfinge's current results are mainly due to a increase in the question's difficulty.

Categories and Subject Descriptors

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; H.2.3 [Database Management]: Languages—Query Languages

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Portuguese, anaphoric reference, parser evaluation, named entity recognition, Wikipedia

1 Introduction

As explained in the QA@CLEF overview [7], 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 respectively in [3, 4, 5]. This year the QA track offered new challenges and most work in Esfinge was related to address those challenges. Figure 1 gives a general overview of the system used this year:

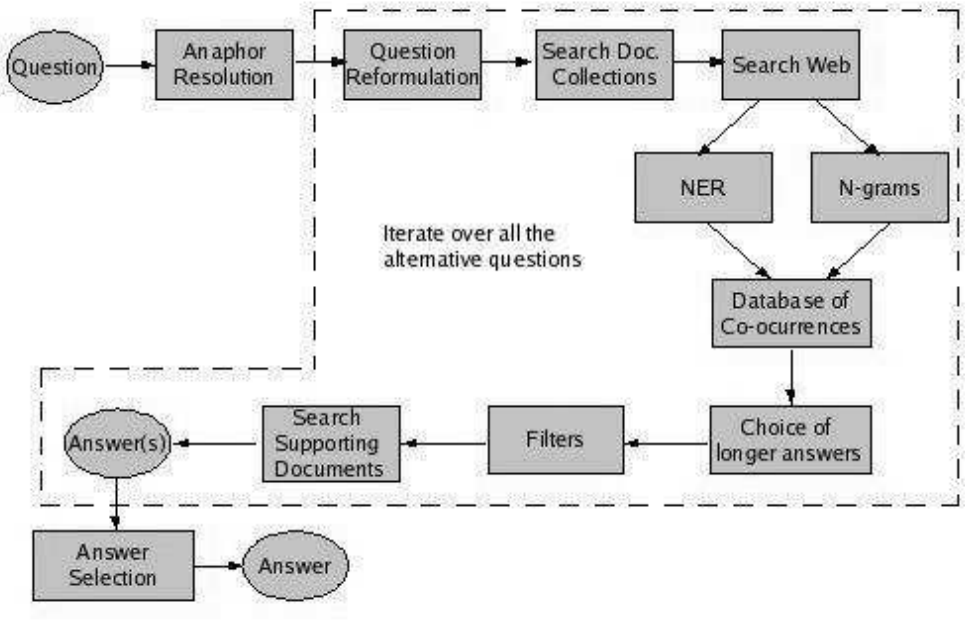

Figure 1: Modules used in Esfinge

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, therefore producing several questions from one.

Then, for each of the alternative questions:

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 the Portuguese Wikipedia (see Section 4 below).

2. If the patterns are not found in the document collections, the system returns the answer NIL (no answer found) and stops. Otherwise, it can proceed by searching the same patterns in the Web (this is optional). Then, all the texts retrieved are analyzed using the named entity recognizer (NER) system SIEMÊS [14] 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 database of co-occurrences BACO [13] and the candidate answers (by ranking order) are

analyzed 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. For example, for the question *O que é M31?* ('What is M31?'), Esfinge found the answer *galáxia* first, but afterwards the candidate answer *galáxia de Andrómeda* ('Andromeda galaxy'), that includes the first and satisfies all remaining criteria. Therefore the answer returned by the system was *galáxia de Andrómeda*, which is actually a better answer.

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

Sets of questions on the same topic are rendered in a more natural way for a human with anaphors and reduced questions, so to deal with such phenomena is a must in QA. We developed a module relying crucially on the PALAVRAS parser [2] to replace anaphoric expressions into fully descriptive expressions (i.e., independently understandable questions).

Given that this question reformulation could be seen as a more general device that produced a set of equivalent questions (see [11] where we introduce the M,N-O,P model), we devised this module as producing an (ordered) set of questions from the input question, that the original Esfinge system should try to answer. Then, another module, described in section 5, would be responsible for choosing among different answers.

We are aware that 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 [1], but we were interested in this model of several question formulations and choice among many answers in general anyway.

3.1 Initial analysis of anaphoric expressions in QA@CLEF questions

We divided the treatment of anaphoricity into five distinct groups, namely: 1) subject pronoun anaphors; 2) other pronominal anaphors; 3) demonstrative anaphors (including both pronominal and nominal anaphora); 4) possessive anaphors and 5) null subject anaphors. After receiving the question set, we discovered that we had forgotten to deal with full sentence anaphora (questions with only a question word), so, although it would have been easy to change the system to produce a new question, we decided not to. Table 1 presents one example of each, together with their distribution on the actual CLEF question sets that included Portuguese as source language.

Table 1: Distribution of the 5 kinds of anaphors catered for

Trigger	Example question	PT-PT	PT-DE	PT-ES	PT-FR	total
subject pronoun	*Quem é o dono delas?*	14	19	6	19	58
personal pronoun	*Quem é que o afundou em 1985?*	1	1	1	0	3
demonstrative pronoun	*Que cinco países da EFTA se juntaram ao EEE quando este entrou em vigor? Quem é que dirige essa agência desde 2005?*	2	0	9	13	24

possessive pronoun	*Qual era o seu verdadeiro nome?*	6	3	4	6	19
null subject	*Quantos habitantes tinha?*	12	1	5	2	20
Total		35	24	25	40	124

So, from Table 1 it can be clearly seen that we did not try to deal with definite description anaphora, nor did we try to deal with implicit anaphora, exemplified respectively by *Quantos lugares tem o estádio?* ('how many places has the stadion') and *Quem é o actor principal?* ('Who is the main actor?').

As to demonstrative nominal anaphora (i.e., cases of noun phrases including the determiners *este*, *esse* or *aquele*), we dealt primarily with 'direct co-reference' [15], although we also tried our luck at some sorts of indirect co-reference that occurred between noun phrases.

But we did not attempt to deal with cases where a significant conversion had to be effected, such as an event into a noun, as illustrated by *Quem era o presidente do regime comunista do Afeganistão antes dessa queda?* ('Who was Afghanistan's president before that fall?'), where *queda* ('fall') is anaphorically referring to the previous question *Quando caiu o regime comunista do Afeganistão?* ('When did Afghanistan's communist regime fall?'), or when general nouns imply a specific evaluation of an event, as in *este feito* ('this deed') pointing to *o voo mais longo sem escalas por um avião civil* ('the longest flight ...').

Then, we found out that there was a considerable number of questions (both with definite description and demonstrative anaphor) which referred directly or indirectly to the answer of some previous question: See *Qual o orçamento dessa universidade para 2005?* ('what was the budget of that university?') after a question "Which university...", or *Quem era considerado chefe do grupo?* ('Who was considered the group's leader?') with *grupo* refering back to *FPLP* in *O que é a FPLP?* ('What is FPLP?'). In fact, to be able to make sense of this last question one would need to know the previous answer,[1] while we were simply modelling anaphoric reference to the previous question(s). Let us remark that, if we were to make use of the answers to the previous questions instead, this would require an altogether different architecture for Esfinge, and we would need a much better performing system to properly test such an architecture anyway.

For another example also needing the answer in order to be fully rephrased, see the sequence of questions *Em que famosa obra aparece o cometa? Que batalha aparece nessa obra?* ('In which famous work does the comet appear? Which battle is described in that work?'). Note that even a human being, provided he was ignorant of the answer to the first question, would not be able to do better than rephrase the second question as something like "Which battle appears in a famous work that mentions the Halley comet?", which seems to be a forbidden kind of question so far in CLEF's history.

3.2 Invoking PALAVRAS

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 initially in [9] and further at the AC/DC project website[2].

For the particular issue here, we were basically interested in identifying the argument phrases, because of our hypothesis that most anaphoric related antecedents would be major constituents. Unfortunately, we found out that the parser was not at all optimized for questions, which, to be fair, constitute a negligible percentage of the input in most kinds of Portuguese text, and

[1] Of course a knowledge-poor approach of treating any set of capitals as an organization could have been used as a heuristic in this case, but our efforts this year were on the development of a robust module for the cases above, which did not even include definite description anaphor.

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

even more so if we restrict our attention to "factoid" (and therefore short) questions.[3] This had as unfortunate consequence that part of our work for anaphor resolution had to be devoted to developing particular parser fixes. Anyway, it should be stressed as well that, to achieve correct anaphoric resolution, correct argument structure was not necessary in every case, and we aimed at a balance between relying on high-precision deep analysis and catering for high-recall shallow approaches. Our solution was then to produce a set of questions with both approaches.

Also, by considering the particular question set it soon became apparent that in some cases syntax alone is not enough to assign the right argument structure. One would have to invoke semantics or world knowledge to decide what is what, cf. the following – completely flawless – Portuguese sentences: *Que sinfonia compôs Beethoven em 1824?* and *Que compositor compôs Vanda em 1875?* (lit. Which symphony-OBJ composed Bethoven-SUBJ in 1824? and Which composer-SUBJ composed Vanda-OBJ in 1875?), where *Beethoven* is the subject of the former and *Vanda* the object of the latter.

Actually, we can state more generally that the simpler the questions, the less syntax is going to help. For example, so far we could make no use of the distinction between subject predicative and subject within QA.

So, given that the subject-object ambiguity in questions seemed to be a general property of Portuguese (rendering both semantic interpretations syntactically admissible), we decided to systematically allow for both possibilities (therefore creating the other syntactic alternative as well), hoping that at least one of them would provide a correct analysis that would then be successfully employed for anaphor resolution.

The result of invoking PALAVRAS gives us then the following information for each original question: the anaphoric element and the phrase it is included in, and a list of possible candidates, formed by all arguments mentioned within the same topic that include a proper name, later augmented by adjuncts with the same property, and finally by all proper names and dates as well.

3.3 Replacing anaphors with their antecedents

As explained in the previous section, we followed a (semantic) knowledge-poor approach, basically identifying (argument) noun phrases, and proper nouns and dates, and had all these as candidates for possible replacements in further questions if anaphoric expressions turned up. It is generally not possible to know, however, whether the anaphoric expression refers to the full noun phrase, or just its head, or even another part of it, as the examples in Table 2 eloquently show, so we chose to produce as many rephrasings as possible.

Table 2: Multiple candidates for the anaphoric expressions

Multiple candidates	Questions with anaphoric expressions
Roma; as sete colinas de Roma;	*Qual é a mais pequena* ***delas****?*
o período de gestação de ocapi;	*Qual o* ***seu*** *peso?*
os primeiros nomes dos dois irmãos Piccard; Piccard;	*Qual* ***deles*** *descobriu o urânio 235?*
Torre dos Clérigos; o estilo arquitectónico da Torre dos Clérigos;	*Qual a altura* ***dela****?*
Steffi Graf, o pai de Steffi Graf,	*Contra quem é que* ***ela*** *não jogou nas meias-finais de Roland Garros em 1994?*
Goethe; Sociedade das Sextas; Imperador José II; a Sociedade das Sextas; a coroação do Imperador José II;	*Em que ano nasceu* ***ele****?*

[3] For lack of space, and also due to the complexity of the particular evaluation task, we leave parser evaluation results to a future paper.

These examples show that there is no simple way to predict which actual antecedent will be felicitous just by looking at the form of the questions or of the candidate antecedents. Of course, number or gender agreement might solve the particular cases of 1, 4 and 5, but this would in addition require extensive information on gender and number at a gazetteer level to be included in the parser, and we chose not to use it.

3.4 Performance of the anaphor resolution module

We assessed the performance of the anaphoric resolution module considering as successful the cases where at least one question reformulation (involving anaphor resolution or its attempt) was able to capture the intended full meaning. According to Mitkov [8], there are two ways of describing the performance of an anaphor resolution module: evaluation of the algorithm (assuming perfect parse) and evaluation of the (real) system (in which parser errors can limit or substantially reduce the overall performance). Table 3 provides here simply the system evaluation, indicating the cases where anaphor resolution was attempted (correctly detected plus spurious) and the case where it actually occurred (from manual inspection), for the five situations addressed by the module. Note that there are two questions in the material where two different ways of resolving anaphors give equally good results. That sentence was counted twice in Table 1 but only once here.

Table 3: Performance of the anaphor resolution module in reality (and given perfect parses) for the 122 questions

	Number of questions	Correctly detected	Spurious	Undetected	Correctly resolved	Accuracy (resolved/all)
PT-PT	34	33	0	1	27	27/34 (79%)
PT-ES	25	24	2	1	14	14/27 (52%)
PT-DE	23	21	5	2	20	20/28 (71%)
PT-FR	40	38	0	2	31	31/40 (78%)
Total	122	116	7	5	92	92/129 (71%)

Two comments should be made regarding Table 3 even if the anaphoric cases were outside the range of our algorithm, provided they were detected and attempted to be resolved, they were included in the table. This accounts for the high number of "answer dependent questions" with a demonstrative determiner that were included in the Spanish material and consequently harmed performance. Also, we considered a rephrasing correct only when the resulting Portuguese sentence was flawless. Thus rephrasings such as **Qual era o de lo Corbusier verdadeiro nome?* in which reference is well resolved but generation of the new sentence is not, were considered incorrect.

3.5 Entities into patterns

A by-product of this 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.

4 Searching Wikipedia

This year the use of Wikipedia presented a new challenge for Esfinge. We chose to implement the access to Wikipedia source in a different way than the one used to access the newspaper collections, for fearing that the size of the text involved would make the current methods prohibitively slow: the initial size of the downloaded articles was about 5.4G, predicting a size of the raw text around 1GB. So, instead of compiling the text as a CQP corpus [6], we stored the Wikipedia source texts in a MySQL database, in a process fairly similar to the one used in BACO, making use of MySQL indexing capabilities to allow faster queries on the collection. Although the raw text size only amounted to 395 MB in the end, this made it possible for us to test other techniques without worrying about query times.

4.1 From HTML to SQL

We used the Wikipedia HTML dump provided by CLEF to create the questions in order to avoid inconsistency with other dumps. The downside of using the aforementioned HTML dump was that we had to process the HTML directly (even though other options would require similar processing of an XML document). Nonetheless, the HTML was well structured and exclusion of unnecessary contents such as the Wikipedia menus was fairly easy. After removing the unwanted parts from the HTML we merely used the `HTML::TokeParser::Simple` Perl package for processing the HTML into text, and `LINGUA::PT::NLP` to split the text into sentences, the smallest information unit we considered.

4.2 Structuring the text

As done in BACO, we created sets of several sentences instead of storing the complete text of an article all together. This seemed a good option considering that each set of sentences would have a higher relevance to the questions posed than a full text would have and these same sets could later be used as the corresponding support texts to the answers. In order to keep the context of the sentences, we repeated information by intercalating the sentences, instead of simply grouping consecutive sentences. Table 4 illustrates this.

Table 4: Storing sentence blocks in MySQL

BLOCK id	Article id	Content
1235	324	SENTENCE N SENTENCE N+1 SENTENCE N+2
1236	324	SENTENCE N+1 SENTENCE N+2 SENTENCE N+3
1237	324	SENTENCE N+2 SENTENCE N+3 SENTENCE N+4

To improve the access time to the data, we indexed the text information. By default MySQL only indexes words with 4 or more characters. As a precaution we changed this minimum value to 3 in order to prevent some words from being ignored. This means that words used in patterns with less than 3 characters were not searchable. We assumed that the impact would not be significant in the final answers, but this has still to be confirmed with a further analysis of the results.

The index creation took only a few hours.

4.3 Searching text patterns

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 had an interface to MySQL that assessed rarity of words in BACO. (See [5] for more details).

Esfinge generates several text patterns from the given question. Each one will then be used to search within the collections. While Esfinge catered for CQP patterns that can be directly applied to the CHAVE collections [12], we had to create corresponding patterns for the MySQL function `Match Against`, which is the method employed for the format of the Wikipedia material.

Table 5: Sample patterns for Wikipedia search from Esfinge Web patterns

Que país declarou a independência em 1291?	
Initial Esfinge expression	**MySQL search expression**
"a independência em 1291" "país declarou" "independência em 1291" "país declarou a" "país declarou a independência em 1291" "declarou a independência em 1291" país país declarou a independência em 1291	+"a independência em 1291" +"país declarou" +"independência em 1291" +"país declarou a" +"país declarou a independência em 1291" +"declarou a independência em 1291" +país +"1291" +país +declarou +a +independência +em

Although the method used in MySQL provides a range of values and options that could be easily applied and experimented with, for this year we had no time to experiment with different solutions and therefore only attempted to obtain results as similar as possible to the ones returned by the previous system. Each search expression is used to obtain results from the colections, independently of word order. While in CQP we make several queries from one expression and later join the results, in MySQL this is done in one single query. This way, the expression"*navegação cabotagem*" succeeds in finding the following sentence: *A **cabotagem** se contrapõe à **navegação** de longo curso, ou seja, aquela realizada entre portos de diferentes nações.*

Each text pattern will be matched against the collections and retrieve possible candidates sets of sentences which will be further analyzed to see if they can provide an answer.

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.

As we had only two possible runs to send, we chose to use the `Answer selection` module also to merge the results of individual runs. We chose to test merging all runs that used the same kind of search patterns (on the one hand, the original Esfinge regular expressions over the questions, on the other hand the search patterns created using PALAVRAS, mentioned in section 3.5 above), as displayed in Table 6.

Table 6: Submitted runs

Using Esfinge patterns from regular expressions			Using PALAVRAS patterns		
Web + Newspapers + Wiki	Newspapers + Wiki	Web + Newspapers	Web + Newspapers + Wiki	Newspapers + Wiki	Web + Newspapapers
esfi0701PTPT.xml			esfi0702PTPT.xml		

To merge the results we took into consideration the following (avowedly very simple) aspects:

1. the number of times (occ.) a certain answer was found in all runs (in a similar way as Esfinge does with the candidate answers),
2. the relevance of the support text to the question asked, WSQ, computed as the number of times that the words (with 3 or more characters) in the question occurred in the support text,

and we used the sum of both counts as our final score.

Table 7: Illustrating the choice among results from different runs.
WSQ stands for number of words in support/question

Question	Answer	#occ.	WSQ	Total
Quando foi fundado o Nacional da Madeira?	**1910**	1	3	**4**
	8 de dezembro de 1910	1	3	4
	NIL	1	--	1
Quais são as cidades-estado da Alemanha?	Berlin, Bremen	1	2	3
	uma região administrativa	1	1	2
	Berlin, Bremen	2	2	**6**
Quando foi inaugurado o metropolitano de Lisboa?	1959	1	3	4
	1755	1	2	3
	1755	2	2	**6**
Quantas ilhas tem Cabo Verde?	Dez	1	3	4
	Quatro	1	3	4
	10	1	3	**4**

Table 7 provides some examples to illustrate the algorithm, while displaying several situations that could be better handled, such as:

1. numbers in natural language or in numerical form: *dez* and *10* are equivalent, but we did not cater for this (although in the particular example the right answer was chosen);
2. dates could be verified: for example, *8 de dezembro de 1910* could strengthen the *1910* answer or vice-versa, and this independently from the fact thatwe would have to choose between the broadest and safest answer (*1910*) or the most acurate one (*8 de dezembro de 1910*)).

This analysis is not only important for improving the combination of answers in the future. It also proved invaluable as feedback for other modules of Esfinge, namely when measuring the relevance of support texts, or ranking candidate answers.

6 Our participation and additional experiments

As described in the previous section, we submitted two official runs. Both are the result of combining several individual runs that used different combinations of data sources. The only difference between the individual runs combined into the two official runs is in the patterns employed to retrieve relevant documents: the runs combined in esfi071PTPT used simple regular expressions (as

in previous participations in CLEF), whereas the runs combined in esfi072PTPT used as search patterns the main verb and possible entities identified by PALAVRAS. Table 8 shows the results of the official runs, together with their subsequent repetition, after several severe bugs were discovered – unfortunately too late to send to the QA@CLEF organizers. In fact, the runs sent were both incomplete, because Esfinge had no time to process all questions, and inaccurate, because faulty versions of some modules had been employed, severely compromising the significance of the results.

Table 8: Results of the official runs and their correction

Runs	Right Answers	Inexact Answers -	Unsupported Answers
esfi071PTPT	15	4	2
esfi071PTPT corrected	57	8	4
esfi072PTPT	11	2	2
esfi072PTPT corrected	36	7	2

Figure 2 displays the results of the individual runs and also of their combination. In order to evaluate our module that performs choice among several answers, we did a manual choice run as well (choosing manually among the different answers). This is indicated as manual combination vs. automatic combination. 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 (described as 2006A and 2006B), which resulted only in a 3-4% improvement. This allows us to conclude that there was a similar proportion of questions this year that were possible to deal with by Esfinge.

Figure 2: Results of the additional experiments (A- Right answers including NIL; B- Partial Right answers on lists; C-Right NIL answers)

#		Description	Right Answers (all questions)			Unsupported Answers	Inexact Answers -	Inexact Answers +	Right Answers (1st questions in 150 topics)			Total NIL
			A	B	C				A	B	C	
	1	Web+News+Wiki	33	2	6	2	7	0	27	2	5	74
	2	News+Wiki	25	0	6	1	3	0	21	0	5	74
	3	Web+news	24	1	8	4	6	0	19	1	6	107
	4	Automatic 1, 2, 3	31	1	6	3	7	0	27	1	5	74
	5	Manual 1, 2, 3	46	2	--	4	8	0	38	2	--	--
PALAVRAS	6	Web+News+Wiki	35	3	5	1	6	1	28	3	3	67
PALAVRAS	7	News+Wiki	25	2	7	3	7	0	19	2	4	98
PALAVRAS	8	Web+News	28	0	5	1	3	1	21	0	3	67
PALAVRAS	9	Automatic 6, 7, 8	34	2	5	2	6	1	27	2	3	68
PALAVRAS	10	Manual 6, 7,|8	49	3	--	2	8	1	38	3	--	--
11		Automatic 1, 2 3,6, 7, 8	34	1	6	2	6	1	30	1	4	73
12		Manual 1, 2 3,6, 7, 8	61	3	--	5	10	1	48	3	--	--
13		Best Run in 2006	50	--	--	3	7	2	--	---	--	--
14		CLEF2006A	57	--	--	6	10	2	--	---	--	--
15		CLEF2006B	56	--	--	4	7	1	--	---	--	--

7 Discussion of results and further work

Although the comparison of Esfinge results in 2006 and 2007 does not allow us to prove this statement, we believe 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, a lot of questions in the test set had errors, which we reported but were maintained by the organization with the argument that this mirrors realistic input. We think this was not beneficial and blurs the distinction between a well-done evaluation and just a random trial. If a user noticed he had written a typo, he would not blame the system; just repeat the question. So such questions, although realistic, in our opinion should not be included in a test collection.

This year, we concentrated mainly on developing the anaphor resolution module and the module responsible for merging and/or choosing from several alternative answers. While the first module seems to have attained relatively good results given the few cases it was subjected to, the choice algorithm apparently managed to always produce worse results than some of the individual runs it combined, and so it deserves further attention. One idea to be pursued is to give different weights to different sources, and/or combine the individual weights that had been assigned in each individual run, or even – and that would be a considerable change in Esfinge's architecture – do combination before choosing candidates.

The anaphoric resolution module and its evaluation allowed us to identify that different question formulation strategies exist in different languages and in particular that different attention to different phenomena will eventually be required to do CLIR into different target languages, something which supports our previous contention that attention to the contrastive aspects should concern the QA@CLEF community more [10, 12].

Regarding future work, a more detailed error analysis is required to better understand where and why Esfinge currently fails. As the questions are getting more and more difficult each year, to improve its results Esfinge will need to use more intelligent ways to retrieve answers, and so we envisage exploring the use of ontologies and further syntactic analysis in the near future.

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] Núria Bertomeu, Hans Uszkoreit, Anette Frank, Hans-Ulrich Krieger, and Brigitte Jörg. 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.

[2] Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University, 2000.

[3] Luís Costa. First Evaluation of Esfinge - a Question Answering System for Portuguese. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum*, pages 522–533, Heidelberg, Germany, 15-17 September 2004 2005. Springer.

[4] Luís Costa. 20th Century Esfinge (Sphinx) solving the riddles at CLEF 2005. In Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Acessing Multilingual information Repositories:6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Vienna, Austria, September*

2005. Revised Selected papers, pages 467–476. Springer Berlin / Heidelberg, 21-23 September 2005 2006.

[5] Luís Costa. Question answering beyond CLEF document collections. In Carol Peters, Paul Clough, Fredric C. Gey, Douglas W. Oard, Maximilian Stempfhuber, Bernardo Magnini, Maarten de Rijke, and Julio Gonzalo, editors, *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September 2006. Revised Selected papers*, Lecture Notes in Computer Science. Springer, Berlin / Heidelberg, 2007.

[6] Stefan Evert. The CQP Query Language Tutorial (CWB version 2.2.b90). Technical report, University of Stuttgart, 10 July 2005.

[7] Bernardo Magnini et al. Overview of the CLEF 2007 Multilingual Question answering track. In This volume.

[8] Ruslan Mitkov. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium*, pages 96–107, 2000. Keynote speech.

[9] Diana Santos and Eckhard Bick. Providing Internet access to Portuguese corpora: the AC/DC project. In Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis, and Gregory Stainhauer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 205–210, 31 May-2 June 2000.

[10] Diana Santos and Nuno Cardoso. Portuguese at CLEF 2005: Reflections and Challenges. In Carol Peters, editor, *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop*, 21-23 September 2005.

[11] Diana Santos and Luis Costa. QolA: fostering collaboration within QA. In Carol Peters, Paul Clough, Fredric C. Gey, Douglas W. Oard, Maximilian Stempfhuber, Bernardo Magnini, Maarten de Rijke, and Julio Gonzalo, editors, *7th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, pages 569–578. Berlin / Heidelberg: Springer, September 2006 2007. Revised Selected papers.

[12] Diana Santos and Paulo Rocha. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum*, pages 821–832, Heidelberg, Germany, 15-17 September 2004 2005. Springer.

[13] Luís Sarmento. BACO - A large database of text and co-occurrences. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik, and Daniel Tapias, editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1787–1790, 22-28 May 2006.

[14] Luís Sarmento. SIEMÊS - A Named Entity Recognizer for Portuguese Relying on Similarity Rules. In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira, and Maria Carmelita Dias, editors, *7th Workshop on Computational Processing of Written and Spoken Language*, pages 90–99. Springer, 13-17 de Maio 2006.

[15] Renata Vieira, Susanne Salmon-Alt, and Caroline Gasperin. Coreference and Anaphoric Relations of Demonstrative Noun Phrases in Multilingual Corpus. In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing: Linguistic, Cognitive and Computation Modelling*, pages 385–401, Amsterdam/Philadelphia, 2005. John Benjamins Publishing Company.