20th Century Esfinge (Sphinx) solving the riddles at CLEF 2005

Luís Costa

Linguateca at SINTEF ICT
Pb 124 Blindern, 0314 Oslo, Norway
luis.costa at sintef.no

Abstract. Esfinge is a general domain Portuguese question answering system. It tries to apply simple techniques to large amounts of text. Esfinge participated last year in the monolingual QA track, but the results were compromised by several basic errors. This year, participation was intended to correct the basic errors of last year and work for the first time in the multilingual QA track.

1 Esfinge overview

The sphinx in the Egyptian/Greek mythology was a demon of destruction that sat outside Thebes and asked riddles to all passers-by. She strangled all the people unable to answer [1], but the times have changed and now Esfinge has to answer questions herself. Fortunately, CLEF's organization is much more benevolent when analysing the results of the QA task. performance

Esfinge (http://acdc.linguateca.pt/Esfinge/) is a question answering system developed for the Portuguese which is based on the architecture proposed by Eric Brill [2]. Brill suggests that it is possible to get state of the art results, applying simple techniques to large quantities of data.

Esfinge starts by converting a question into patterns of plausible answers. These patterns are queried in several text collections (CLEF text collections and the Web) to obtain snippets of text where the answers are likely to be found.

Then, the system harvests these snippets for word N-grams. The N-grams will be later ranked according to their frequency, length and the patterns used to recover the snippets where the N-grams were found (these patterns are scored a priori). Several simple techniques are used to discard or enhance the score of each of the N-grams. Finally the answer will be the top ranked N-gram or NIL if neither of the N-grams passes all the filters.

2 Strategies for CLEF 2005

During last year participation, several problems compromised the results. The main objectives for this year were to correct these problems, and to participate in the multilingual tasks.

This year, in addition to the European Portuguese text collection (Público), the organization also provided a Brazilian Portuguese collection (Folha). This new collection helped Esfinge, since one of the problems encountered last year was precisely that the document collection only had texts written in the European variant and some of the answers discovered by the system were in the Brazilian variant, therefore difficult to justify [3].

2.1 Pre-processing

IMS Corpus Workbench [4] was used again to encode the document collections. Each document was divided in sets of three sentences. Last year other text unit sizes were tried (namely 50 contiguous words and one sentence), but the results using three sentence sets were slightly better. The sentence segmentation and tokenization was done using the Perl Module Lingua::PT::PLNbase developed at Linguateca and freely available at CPAN. For the English documents, the sentence segmentation and tokenization programs used by DISPARA in the COMPARA project [5] were used.

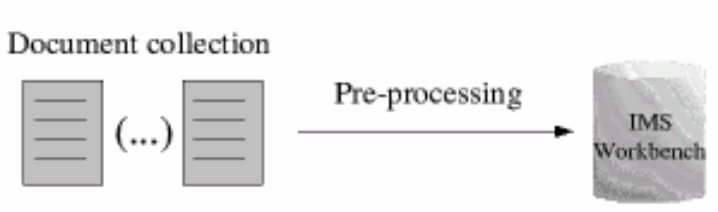

Figure 1

2.2 PT-PT monolingual task

Two different strategies were tested. In the first one, the system searched the answers in the Web and used the CLEF document collection to confirm these answers (Run 1). In the second one, it searched the answers in the CLEF document collection only (Run 2).

Run 1

This experiment used the strategy described in another paper by Brill [6]: answers are searched in the Web, and then the system tries to find documents in the document collection supporting those answers.

For each question in the QA track, Esfinge performed the following tasks:

Question reformulation. The question is submitted to the question reformulation module. This module uses a pattern file that associates patterns of questions with patterns of plausible answers. The result is a set of pairs (answer pattern, score). Some patterns were added this year to the patterns file, based on last year's questions. The following pattern is one of the patterns included in that file:

Onde ([^\s?]) ([^?]*)\??/"$2 $1"/20*

It means that for a question including the word *Onde* (Where*)*, followed by some words, a possible pattern for an answer will be the words following the one immediately after *Onde*, followed by the word after *Onde* in a phrase pattern.

As an example, take the question *Onde fica Lillehammer?* (Where is Lillehammer located?). This generates the pattern *Lillehamer fica* with a score of 20, that can be used to search for documents containing an answer to the question.

Passage extraction. The patterns obtained in the previous module are submitted to Google. Then, the system extracts the document snippets $\{S_1, S_2 \ldots S_n\}$ from Google's results pages.

It was detected in the experiments made with the system that certain types of sites may compromise the quality of the returned answers. To overcome this problem it was created a list of address patterns which are not to be considered (the system does not consider documents stored in addresses that match these patterns). This list includes patterns such as *blog*, *humor*, *piadas* (jokes). These patterns were created manually, but in the future it may be rewarding to use more complex techniques to classify web pages [7].

Another improvement over last year experiment was that if no documents are recovered from the Web, the system tries to recover them from CLEF's document collection. When searching in the document collection, the stop-words without context are discarded. For example in the query *"o" "ditador" "cubano" "antes" "da" "revolução"* (the Cuban dictator before the revolution), the words *o* and *da* are discarded while in the query *"o ditador cubano antes da revolução"* (phrase pattern) they are not discarded. Last year the 22 most frequent words in the CETEMPúblico corpus [8] were discarded. This year in addition to those, some other words were discarded. The choice of these words was the result of the tests performed with the system. Some examples are *chama* (is called), *fica* (is located), *país* (country) and *se situa* (is). One may find these words in questions, but using them in the search pattern may increase the difficulty to find documents containing its answers. An example is the question *Com que país faz fronteira a Coreia do Norte?* (What country does North Korea border on?). It is more likely to find sentences like *A Coreia do Norte faz fronteira com a China* (North Korea borders with China) than sentences including the word *país*.

When the system neither recovers documents from the Web, nor from CLEF's document collection, one last try is made by stemming some words in the search patterns. The system uses the morphological analyser *jspell*

[9] to check the PoS of the various words in each query. Then the words classified as common nouns, adjectives, verbs and numbers are stemmed using the module Lingua::PT::Stemmer freely available at CPAN, implementing a Portuguese stemming algorithm proposed by Moreira & Huyck [10]. This provides the system with more general search patterns that will be used to search documents in the document collection.

If documents are retrieved using any of the previous techniques, at the end of this stage the system has a set of document passages {P1, P2 … Pn} hopefully containing answers to the question. If no documents are retrieved, the system stops here and returns the answer NIL (no answer found).

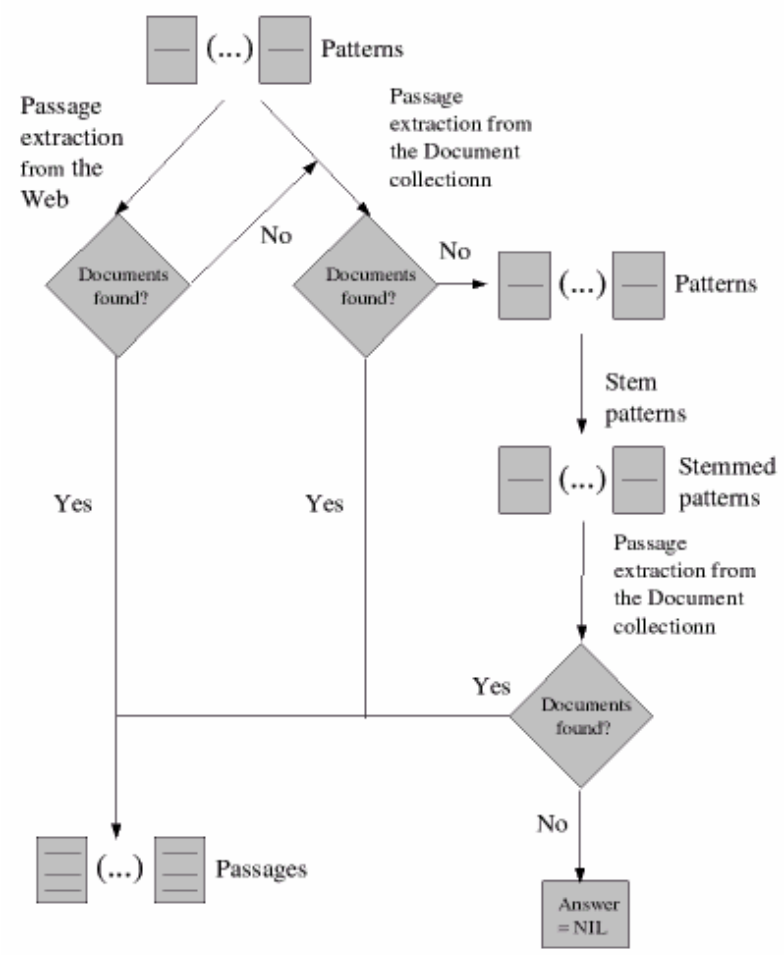

Figure 2

N-grams harvesting. The distribution of word N-grams (from length 1 to length 3) of the first 100 document excerpts recovered on the previous module is computed. The system uses the Ngram Statistics Package (NSP) [11] for that purpose.

Then, the word N-grams are ordered using the following formula:

N-gram score = $\sum$ (F * S * L), through the first 100 snippets resulting from the web search where:

F = N-gram frequency
S = Score of the search pattern which recovered the document
L = N-gram length

At the end of this stage, the system has an ordered set of possible answers {A1, A2 … An}.

Named entity recognition/classification in the N-grams. This module was developed for this year participation, hoping that the use of a named entity recognition (NER) system could improve the results (at least for some types of questions).

An extra motivation for using a NER system was the HAREM (Evaluation Contest of Named Entity Recognition Systems for Portuguese) [12]. This event boosted the development or improvement of already existent NER systems for Portuguese. One of the participants was SIEMES [13] which was developed in the Linguateca node located in Porto, and obtained the best recall among all the systems participating in HAREM.

SIEMES detects and classifies named entities in a wide range of categories. Esfinge used a sub-set of these categories: Human, Country, Settlement (includes cities, villages, etc), Geographical Locations (locations with no political entailment, like for example Africa), Date and Quantity.

Esfinge uses a pattern file that associates patterns of questions with the type of expected result. The following pattern is included in that file:

Quant(o|a)s./VALOR TIPO="QUANTIDADE*

This pattern means that a question starting with Quantos (how many – masculine form) or Quantas (how many – feminine form) should have a QUANTIDADE (quantity) type answer.

What the system does in this module is to check whether the question matches with any of the patterns in the "question pattern"/"answer type" file. If it does, the 200 best scored word N-grams are submitted to SIEMES. Then the results returned by SIEMES are analysed to check whether the NER system recognizes named entities classified as one of the desired types. If such named entities are recognized, their ranking in the list of possible answers will be enhanced.

The NER system is used in the "Who" questions in a slightly different way. First it is used to check whether there is a person in the question and if that happens, the NER system is not invoked on the candidate answers (example: *Who is Fidel Ramos?*). There are some exceptions to this rule however and some special patterns to deal with them too (example: *Who is John Lennon's widow?*). When there is not a person in the question, the NER system is always invoked to find instances of persons for the Who questions.

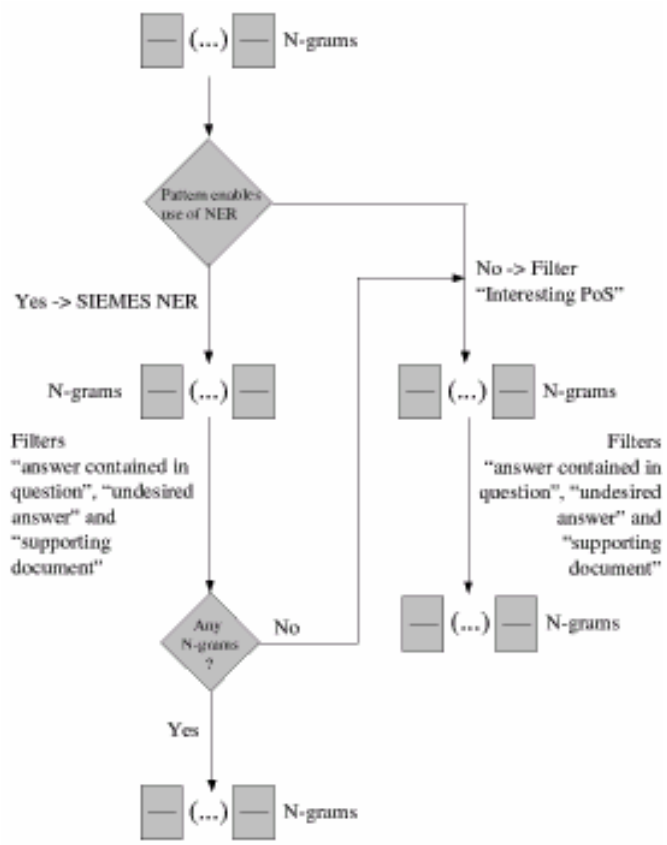

Figure 3

N-gram filtering. In this module the list of possible answers is submitted to a set of filters (by ranking order), namely:

- A filter that discards words contained in the questions. Ex: the answer *Satriani* is not desired for the question *Quem é Joe Satriani?* (Who is Joe Satriani?) and should be discarded.

- A filter that discards answers contained in a list of 'undesired answers'. This list was built with the help of Esfinge's log file. The frequency list of all the solutions provided by Esfinge to the 2004 CLEF QA track questions was computed (not only the best answer, but all the answers that managed to go through all system's filters). With this frequency list and some common sense, the list of 'undesired answers' was built. The words in this list are frequent words that do not really answer questions in isolation (like *pessoas*/persons, *nova*/new, *lugar*/place, *grandes*/big, *exemplo*/example). Later some other answers were added to this list, as a result of the tests performed with the system. The list includes now 92 entries.

- A filter that uses the morphologic analyser *jspell* [9] to check the PoS of the various tokens in each answer. This filter is only used if the system could not predict the type of answer for the question (using the "question pattern"/"answer type" file) or if SIEMES was not able to find any answer of the desired type. Jspell returns a set of possible PoS tags for each token. Esfinge considers some PoS as "interesting": adjectives (adj), common nouns (nc), numbers (card) and proper nouns (np). All answers whose first and final token are not classified as one of these "interesting" PoS are discarded.

- A filter that checks whether the system can find a document supporting the answer in the collection. This filter is only used if the system retrieved documents from the Web. When the system cannot retrieve documents from the Web, it retrieves them from CLEF's document collection, and since the N-grams are extracted from these documents there is no need for this filter. It searches the document collection for documents containing both the candidate answer and a pattern obtained from the question reformulation module.

N-gram composition. The motivation to use this very simple module arose from the analysis of last year's results and some additional tests performed in the system. Sometimes the answers returned by the system were fragments of the right answers. To minimize this problem, a very simple composition algorithm was implemented this year. When an answer passes all the filters in the previous module, the system does not return that answer immediately and stops like in last year. Instead it checks whether there are more candidate answers containing the answer which was found. Each of these candidate answers are submitted to the filters described in the previous module and if one of them succeeds to pass all the filters, this candidate answer becomes the new answer to be returned as result.

Final answer. The final answer is the candidate answer with the highest score in the set of candidate answers which are not discarded by any of the filters described above. If all the answers are discarded by the filters, then the final answer is NIL (meaning that the system is not able to find an answer in the document collection).

Run 2

The difference in this run was that the Web was not used as a resource. The answers were only searched in CLEF's document collection. Consequently, another difference to the algorithm used for the first run was that it was not necessary to check whether there was a document in the collection supporting the answers found since the document collection was the only source used to find them.

2.3 EN-PT multilingual task

In this experiment the questions were translated using the module Lingua::PT::Translate freely available at CPAN. This module provides an easy interface to Altavista's Babelfish translating tool.

After the translation this experiment followed the algorithm described for the PT-PT monolingual task in run 1 (the run which seemed to have the best results).

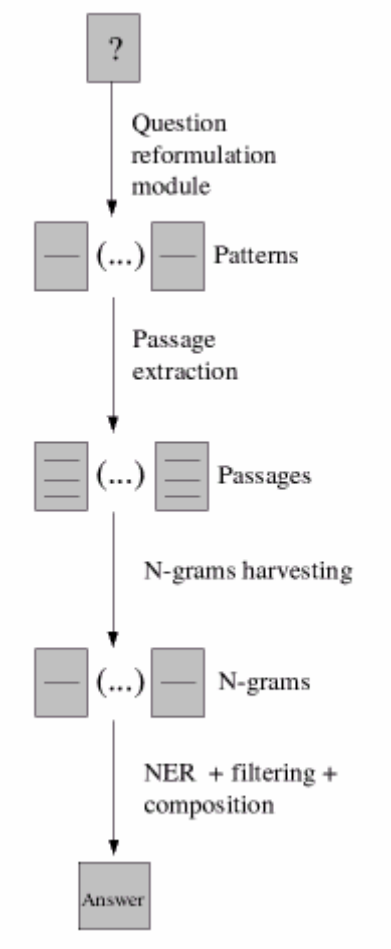

Figure 4

3 Results

The results sent to the organization are presented and discussed in this section together with the error analysis performed for one of the runs and some considerations about CLEF 2005 set of questions motivated by this error analysis. To create the tables, the question set was divided in categories that intended to check how well the various strategies used by Esfinge perform. For example the category "People" includes all questions where the system expects to have the name of a person as answer such as the "Who" questions, in which the NER system is invoked to find names of persons in the recovered documents. This assumption is usually correct, but there are some exceptions however. An example is the question *Who was Barings taken over by?* in which the answer is not a person, but a bank. Other interesting categories are "Places", "Quantities" and "Dates" where the NER system is also used to find instances of those categories in the recovered texts. The category "What is the name of X" does not include some matching patterns, in which it is easy to infer that the answer will be of type person. *What is Nick Leeson's wife's name?* is a good example. The same applies for the categories "Name X" and "Which X" in which some questions are not included and are instead placed in another category because the type of answer is easy to infer. Examples of these kinds of questions are *Name a city with 650,000 inhabitants* and *Which country is Alexandria in?*

3.1 PT-PT monolingual task

	# questions	# Run 1	% Run 1	# Run 2	% Run 2	#questions 2004	# Esfinge's best 2004	% Esfinge's best 2004
People	47	11	23%	15	32%	43	8	19%
"(Que|Qual) X" - "Which X"	36	9	25%	5	14%	42	7	17%
Place	33	9	27%	7	21%	41	7	17%
"Quem (é|foi|era) <HUM>" - "Who (is|was) <HUM>"	27	6	22%	6	22%	17	2	12%
Quantity	18	4	22%	3	17%	23	4	17%
Date	15	3	20%	5	33%	15	0	0%
"Que é X" – "What is X"	15	2	13%	0	0%	15	1	7%
Como se chama|chamou|chamava X – What is X called	5	4	80%	2	40%	0	0	0%
Mencione/Indique/Nomeie X - Name X	4	0	0%	0	0%	3	1	33%
Total	200	48[1]	24%	43	22%	199	30	15%

Table 1. Results by type of question in the PT-PT monolingual task

From table 1 it is possible to conclude that the runs submitted for the Portuguese source/Portuguese target task obtain similar results. The run that used the Web (Run 1) got slightly better results, as last year. One can also see that the results in Run 1 are more homogenous than the ones in the second run. Some results are consistently bad, like definitions not involving people (What is X) and not obvious naming (Name X), but that is not surprising since Esfinge does not have special features to deal with definitions. The results of the second run for the questions of type "People" and "Date" are better both comparing to the other types of questions and to the same type of questions in the first run.

Comparing with last year's results (right columns in the table), one can see that the results improved consistently in almost all types of questions.

	# questions	# Run 1 and Run 2	%	# Run 1 and not(Run 2)	%	# Run 2 and not(Run 1)	%	# Run 1 or Run 2	%
People	47	8	17%	3	6%	7	15%	18	38%
"(Que|Qual) X" - "Which X"	36	4	11%	5	14%	0	0%	9	25%
Place	33	5	15%	4	12%	2	6%	11	33%
"Quem (é|foi|era) <HUM>" - "Who (is|was) <HUM>"	27	2	7%	4	15%	4	15%	10	37%
Quantity	18	3	17%	1	6%	0	0%	4	22%
Date	15	2	13%	1	7%	3	20%	6	40%
"Que é X" – "What is X"	15	0	0%	2	13%	0	0%	2	13%
Como se chama|chamou|chamava X – What is X called	5	2	40%	2	40%	0	0%	4	80%
Mencione/Indique/Nomeie X - Name X	4	0	0%	0	0%	0	0%	0	0%
Totals	200	26	13%	22	11%	16	8%	64	32%

Table 2. Combined results

[1] The official result is 46 right answers, but during the evaluation of the results I found two more right answers.

Table 2 shows the number of questions with right answers in both runs (Run 1 and Run 2), the number of questions with right answers only on the first run (Run 1 and not(Run 2)), the number of questions with right answers only on the second run (Run 2 and not(Run 1)) and the number of questions with a right answer in at least one of the runs (Run 1 or Run 2).

One can observe that the two runs perform better with different types of questions, which suggests that both of the strategies used are still worthwhile to experiment and study.

Problem	# Wrong answers
No documents recovered in the Document collection	42
Answer scoring algorithm	30
No documents recovered containing the answer	25
No documents recovered in the Web	23
Error in tokenization	19
Filter "documents supporting answer"	15
Answer length >3	13
Problems with the NER system	11
Missing patterns in the file "question pattern"/"answer type"	9

Table 3. Causes for wrong answers

The system's log file was used to investigate the causes for the wrong answers. The system registers in this file all the analysed word N-grams for each of the questions. When word N-grams are rejected by some of the filters, this information is also recorded in the log file.

In Table 3, a detailed error analysis for the first run is provided. For some of the questions, it was possible to detect more than one reason for failure. In these cases, both reasons were counted.

From this evaluation, it is possible to create sets of questions with the same type of problems that can be used to debug and improve the system.

3.2 EN-PT multilingual task

	# questions	# Right answers	% Right answers
People	47	6	13%
"(Que|Qual) X" - "Which X"	36	6	17%
Places	33	2	6%
"Quem (é|foi|era) <HUM>" - "Who (is|was) <HUM>"	27	6	22%
Quantities	18	1	6%
Dates	15	2	13%
"Que é X" – "What is X"	15	0	0%
Como se chama|chamou|chamava X – What is X called	5	2	40%
Mencione/Indique/Nomeie X - Name X	4	0	0%
Totais	200	25	13%

Table 4. Results by type of question in the EN-PT multilingual task

From the results in Table 4 it is possible to conclude that most of the questions with right answers are the ones where the NER system was not used (14 out of 25). However, an error analysis similar to the one performed for the PT-PT task will be needed to take more solid conclusions.

3.3 Some considerations about CLEF 2005 set of questions

The error analysis is not only useful to find the reasons motivating system errors. Here and there one is confronted with some interesting cases. I will describe two of them.

The question *Who is Josef Paul Kleihues?* doesn't have an answer in the document collection according to the organization, but is this really true? There is a document with the following text (freely translated from the Portuguese original):

People from Galicia like good architecture. In Santiago de Compostela, besides the "Centro Galego de Arte Contemporânea" designed by Siza Vieira, it was built in the historical center a gym designed by the german Josef Paul Kleihues.

One of Esfinge's runs returned the answer Arquitectura (*architecture*) giving as support the text from where the previous excerpt was extracted. One may question which answer would be more useful for a hypothetical user.: NIL or the answer provided by Esfinge?

I found another curious example in the question *Which **was** the largest Italian party?*. On one of the runs Esfinge returned the answer *Força Itália* supporting it with a document stating that *Força Itália* **is** the largest Italian party (it was true at the time the document was written). The organization considered this answer wrong, however, because they wanted an Italian party that was the largest in the past, but was no longer the largest.

In my opinion the answer provided by the system was acceptable, because the question is being asked in 2005, so one can ask *which **was** the largest Italian party,* and one can support an answer with a document from 1994 saying that *the largest Italian party **is** X.*

Although I can understand the point of view of the organization, I think that this kind of question is confusing and polemic even for humans, therefore not particularly useful to evaluate Q&A systems.

4 Additional experiments

The error analysis (condensed on table 3) provided an insight on the problems affecting the system's performance.

Some effort was invested in the problems that seemed easier to solve. Namely on the "Error in tokenization", "Problems with the NER system" and "Missing patterns in the file question pattern/answer type". The results of the system after this improvement using the same strategy as in Run 1 are presented in table 5. On that table it is also possible to check how each part of the system helps global performance: the results obtained either without using the NER system or without using the morphological analyser are presented. One can see that (in different types of questions) both this components are helping the system.

	# questions	# Run 3	% Run 3	# No NER	% No NER	# No PoS filtering	% No PoS filtering
People	47	14	30%	9	19%	13	28%
"(Que|Qual) X" - "Which X"	36	11	31%	--	--	7	19%
Places	33	10	30%	9	27%	12	36%
"Quem (é|foi|era) <HUM>" - "Who (is|was) <HUM>"	27	7	26%	--	--	3	11%
Quantities	18	3	17%	1	6%	3	17%
Dates	15	8	53%	3	20%	6	40%
"Que é" – "What is"	15	4	27%	--	--	2	13%
Como se chama|chamou|chamava	5	3	60%	--	--	2	40%

Mencione/Indique/Nomeie - Name	4	1	25%	--	--	0	0%
Totais	200	61	31%	48	24%	48	24%

Table 5. Results in the PT-PT monolingual task after improvements in the system using the first run strategy

Applying the system to the 2004 questions after the improvements and using the same strategy as in Run 1 provides the results presented in table 6. The cause for the better results this year could be the possibility that this year's questions were easier than last year's, but this table shows that the system performs better with last year's questions as well.

	# questions 2004	# Esfinge's best 2004	% Esfinge's best 2004	# Run 4	% Run 4
People	43	8	19%	15	35%
"(Que|Qual) X" - "Which X"	42	7	17%	9	21%
Place	41	7	17%	17	41%
"Quem (é|foi|era) <HUM>" - "Who (is|was) <HUM>"	17	2	12%	1	6%
Quantity	23	4	17%	4	17%
Date	15	0	0%	4	27%
"Que é X" – "What is X"	15	1	7%	4	27%
Mencione/Indique/Nomeie X - Name X	3	1	33%	1	33%
Total	199	30	15%	55	28%

Table 6. Results in the PT-PT task after improvements in the system using the first run strategy on 2004 questions

5 Concluding remarks

The results show that Esfinge improved comparing to last year: the results are better both with this year's and last year's questions. Another conclusion is that the two tested strategies perform better with different types of questions, which suggests that both are still worthwhile to experiment and study further.

The experiments performed to check how each part of the system helps global performance shown that (in different types of questions) both the NER system and the morphological analyser improve the system's performance.

6 Acknowledgements

I thank Diana Santos for reviewing previous versions of this paper, Alberto Simões for the hints on using the Perl Modules "jspell" , "Lingua::PT::PLNbase" and Lingua::PT::Translate, Luís Sarmento, Luís Cabral and Ana Sofia Pinto for supporting the use of the NER system SIEMES and Paul Rayson for supporting the use of CLAWS Web Tagger [14] (it was planned to send a run for the PT-EN multilingual task, but it was not possible to finish it in time to send it to the organization).

This work is financed by the Portuguese Fundação para a Ciência e Tecnologia through grant POSI/PLP/43931/2001, co-financed by POSI.

References

1. Wikipedia: http://en.wikipedia.org/wiki/Sphinx/
2. Brill, E.: Processing Natural Language without Natural Language Processing. In: Gelbukh, A. (ed.): CICLing 2003. LNCS 2588. Springer-Verlag Berlin Heidelberg (2003) pp. 360-9
3. Costa, L.: First Evaluation of Esfinge - a Question Answering System for Portuguese. In Peters C., Clough P., Gonzalo J., Jones G., Kluck M. & Magnini B. (eds.), *Advances in Cross-Language Information Retrieval: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)* (Bath, UK, 15-17 September 2004), Heidelberg, Alemanha: Springer. Lecture Notes in Computer Science (to be published)
4. Christ, O., Schulze, B.M., Hofmann, A. & Koenig, E.: The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. University of Stuttgart, March 8, 1999 (CQP V2.2)
5. Santos D.: DISPARA, a system for distributing parallel corpora on the Web, in Elisabete Ranchhod & Nuno J. Mamede (eds.), *Advances in Natural Language Processing (Third International Conference, PorTAL 2002, Faro, Portugal, June 2002, Proceedings)* , LNAI 2389, Springer, 2002, pp.209-218.
6. Brill, E., Lin, J., Banko, M., Dumais, S. & Ng, A.: Data-Intensive Question Answering. In: Voorhees, E.M. & Harman, D.K. (eds.): Information Technology: The Tenth Text Retrieval Conference, TREC 2001. NIST Special Publication 500-250. pp. 393-400
7. Aires, R., Aluísio, S. & Santos, D.: User-aware page classification in a search engine. In *Proceedings of Stylistic Analysis Of Text For Information Access, SIGIR 2005 Workshop* (Salvador, Bahia, Brasil, 19 de Agosto de 2005).
8. Santos, D. & Rocha, P.: Evaluating CETEMPúblico, a free resource for Portuguese. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (Toulouse, 9-11 July 2001) pp. 442-449
9. Simões, A. M. & Almeida, J.J.: Jspell.pm - um módulo de análise morfológica para uso em Processamento de Linguagem Natural. In: Gonçalves, A. & Correia, C.N. (eds.): Actas do XVII Encontro da Associação Portuguesa de Linguística (APL 2001) (Lisboa, 2-4 Outubro 2001). APL Lisboa (2002) pp. 485-495
10. Orengo, V. M. & Huyck, C.: A Stemming algorithm for the Portuguese Language. In *8th Internacional Symposium on String Processing and Information Retrieval (SPIRE'2001)* (Laguna de San Rafael, Chile, 13-15 de Novembro de 2001), IEEE Computer Society Publications, pp. 183-193.
11. Banerjee, S. & Pedersen, T.: The Design, Implementation, and Use of the Ngram Statistic Package. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (Mexico City, February 2003) pp. 370-381
12. Santos, D. : Relatório da Linguateca de 15 de Maio de 2004 a 14 de Maio de 2005, pp. 16
13. Sarmento, L., Pinto, A. S., Cabral, L.: REPENTINO – A collaborative wide-scope gazetteer for Entity Recognition in Portuguese. (to be published)
14. Rayson, P. & Garside, R. (1998). The CLAWS Web Tagger. ICAME Journal, no. 22. The HIT-centre - Norwegian Computing Centre for the Humanities, Bergen, pp. 121-123