Cooperation between black box and glass box approaches for the evaluation of a question answering system

Martine Hurault-Plantet*, **Laura Monceaux***

*LIMSI-CNRS
Bat. 508, Universite Paris XI, 91 403 Orsay, France
{mhp, monceaux}@limsi.fr

Abstract

For the past three years, the question answering system QALC, currently developed in our team, has been taking part in the Question Answering (QA) track of evaluation campaigns TREC (Text REtrieval Conference). In the QA track, each system is evaluated according to a black box approach: as input, a set of questions, and as output, for each question, five answers ranked with regard to decreasing relevance. A score is then computed with regard to the correctness of the answers. Such an evaluation is attractive for comparing systems to each other, as well as for comparing a system to itself after a modification. However, the capacity for knowing how to improve the system requires another approach: the glass box approach. Indeed, in complex modular systems such as question answering systems, we have to ”enter” inside the system and evaluate each module in order to assess if it reaches the goal that has been set for it, or not. Nevertheless, after modifying a module, we have to apply again the back box approach on the whole system in order to judge the effect of the modifications on the overall result. In this paper, we thus present an evaluation of our system, based both on black box and glass box approaches. We will describe the methods used as well as the results that we obtain.

1. Introduction

For the past three years, the question answering system QALC (Ferret et al., 2001b), currently developed in our team, has been taking part in the evaluation campaigns TREC (Text REtrieval Conference) organized by NIST[1] (National Institute of Standards and Technology). The Question Answering (QA) track of TREC involves searching for answers to a list of questions, within a collection of documents provided by NIST. Questions are factual or encyclopaedic, while documents are newspaper articles.

In the QA track, each system is evaluated according to a black box approach: as input, a set of questions, and as output, for each question, five answers ranked with regard to decreasing relevance. Answers have to be short, less then 50 characters, and have to include the number of the document from which each of them has been retrieved. Human judges decide on the correctness of each answer. Global evaluation of each system is then computed with regard to both this judgment and the ranking of the correct answer among the five answers provided for each question (Voorhees and Tice, 2000).

Such an evaluation is attractive for comparing systems to each other, as well as for comparing a system to itself after a modification. However, the capacity for knowing how to improve the system requires another approach: the glass box approach. Indeed, in complex modular systems such as question answering systems, we have to ”enter” inside the system and evaluate each module in order to assess if it reaches the goal that has been set for it, or not. Nevertheless, after modifying a module, we have to apply again the back box approach on the whole system in order to judge the effect of the modifications on the overall result. Both approaches, black box and glass box, have already been used in the evaluation other modular complex systems such as dialog systems (Simpson and Fraser, 1993).

In this paper, we thus present an evaluation protocol combining the black box and the glass box approaches. In the following section, we describe the QALC system and the methodology of evaluation. We then present the results of the black box evaluation of the whole question corpus and for each question category. Afterwards, we present the evaluation of each module of the system, focusing on the most interesting categories, i.e. those which have some salient feature of size or of behaviour. Finally, we dicuss the main results of the evaluation of the system, and we end with a few concluding remarks.

2. Evaluation framework

2.1. System architecture

The QALC system is made of three main modules, one devoted to the processing of the questions, another one to the corpora, and the last module which extracts the answer from the documents by using the informations collected by the two other modules. Each of these modules includes a number of processes (see Figure 1).

2.1.1. Question processing module

This module includes a question analysis process and a term extractor. The term extractor is based on syntactic patterns which describe compound nouns. The maximal extension of these compounds is produced along with the plausible sub-phrases. All the noun phrases belonging to this maximal extension are also produced.

The analysis of the question reasons about the outputs of a shallow parser in order to extract a number of informations from the question:

- an expected answer type that corresponds to the types of entities which are likely to constitute the answer to this question. The answer type may be a named entity list (for example, Person, Organization, Location-city) or the semantic type which corresponds to an item of the lexical base WordNet (Fellbaum, 1998),

[1] http://trec.nist.gov/

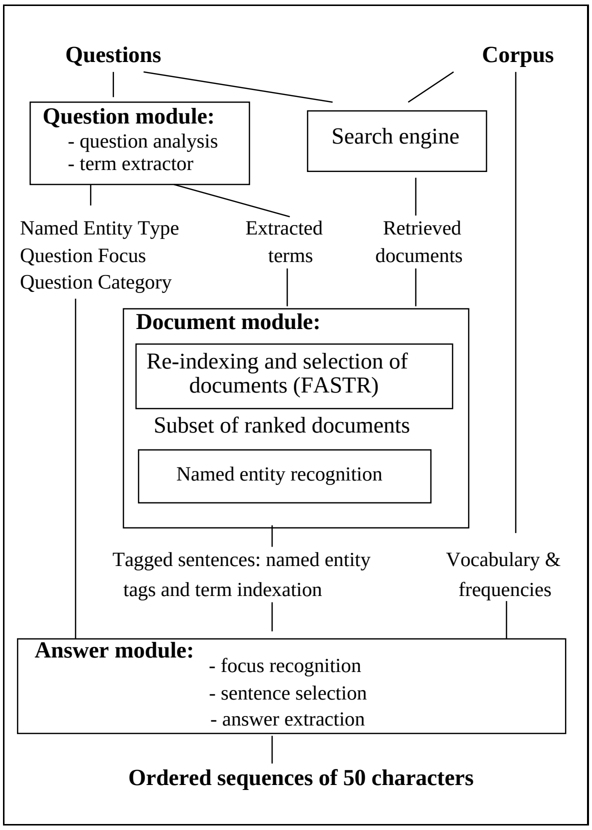

Figure 1: Architecture of QALC system

- a question focus that corresponds to a noun phrase that is likely to be present in the answer. The question module determines the focus, but also the focus-head and its modifiers,
- a question category which corresponds to the syntactic form of the question.

For example, the question module returns for the next question *Who was the first governor of Alaska ?* these different informations:

Question : Who was the first governor of Alaska ?
Category = WhobeNP[2]
Named Entities List = { PERSON }
Focus = the first governor of Alaska
Focus-Head = governor
Focus-Head-Modifiers = ADJ first, COMP Alaska

2.1.2. Document processing module

The QALC system takes as input the top 200 documents from the list of the 1000 best ranked documents retrieved by the search engine of the NIST for the set of questions of the TREC conference evaluation. Such a list is provided at the beginning of the each evaluation campaign. The 200 best documents are re-indexed by Fastr (Jacquemin, 1999), a shallow transformational natural language analyzer which recognizes the occurrences and the variants of the terms produced by the term extraction process. Each occurrence or variant constitutes an index to the document which is ultimately used in the process of document ranking and in the process of question/document pairing. These indexes allow QALC to reorder the documents and entail the selection of a subpart of them (Ferret et al., 2001a). A named entity recognition is then applied on the resulting set of documents.

2.1.3. Answer module

This module relies on two main processes : the sentence selection and the answer extraction. All the data extracted from the questions and the documents by the previous modules are used by a pairing module to measure the similarity between a document sentence and a question.

The answers are then extracted from the more relevant sentences. The extraction process depends on whether the expected answer type is, or is not, a named entity. Indeed, when the answer type is a named entity, the extraction consists of the location of the named entity within the sentence. Thus it mainly relies on the results of the named entity recognition module. On the other hand, when the answer type is not a named entity, the extraction process mainly relies on the recognition of the question focus, as it consists of the recognition of focus-based syntactic answer patterns within the sentence.

The syntactic patterns for answer extraction always include the noun phrase of the focus-head in the sentence and the noun phrase of the answer. Those two elements are usually connected by other elements such as comma, quotation marks, a preposition or even a verb. The only exception occurs when the answer is within the noun phrase of the focus-head. In this case, there is no connecting element between the noun phrase of the focus-head and the noun phrase of the answer in the corresponding syntactical pattern. We distinguished three different pattern structures:

(1) NPfocus Connecting-elements NPanswer
(2) NPanswer Connecting-elements NPfocus
(3) NPanswer-within-NPfocus

Example:
Question: What is the most popular sport in Japan?
Focus = the most popular sport
Focus-Head = sport
Answer : baseball as the nation's most popular sport

In this example, the answer has been extracted through the pattern *NPanswer as NPfocus* from the candidate sentence *Now, it is threatening to dislodge Japan's stodgy baseball as the nation's most popular sport..*

2.2. Evaluation methodology

We carried out an evaluation of our system, based both on black box and glass box approaches. With the aim of defining more precisely the improvements that we need to achieve in our system, we chose to partition the question corpus according to the question categories. Once we achieved this corpus partition, we first apply the black box approach. This evaluation uses the same measure as the one used in TREC, i.e. a score which depends on the cor-

[2] NP means noun phrase, ADJ means adjective, and COMP, complement.

Category	Example	Number of questions	Score
Where	Where is the Holland Tunnel?	27	0.316
When	When did Hawaii become a state?	26	0.280
WhatNPdoNP	What year did the U.S. buy Alaska?	24	0.272
Who	Who discovered x-rays?	46	0.254
WhatbeNPofNP	What is the melting point of copper?	47	0.247
How	How long did Rip Van Winkle sleep?	33	0.192
WhatbeNP	What is acupuncture?	199	0.189
WhatNPbeNP	What precious stone is a form of pure carbon?	47	0.182
WhatNPverbNP	What strait separates North America from Asia?	6	0.167
Which	Which president was unmarried?	10	0.100
WhatdoNP	What does a barometer measure?	22	0.027
Why	Why does the moon turn orange?	4	0.000

Table 1: Evaluation of the overall system for each question category

rectness of the answer and on its rank (Voorhees and Tice, 2000). This score is derived from the mean reciprocal rank of the first five answers. For each question, the first correct answer, among the first five answers, get a mark in reverse proportion to its rank. The score for all the questions is the mean of the question marks. Thus, we obtain an evaluation of the whole system for each subset of questions. This first evaluation enables us to know for which sets of questions the system is effective, but it does not explain why. Therefore, the next step consists of a glass box approach applied to sets of questions, and especially to those that have been previously marked as less effective.

The glass box approach consists of an evaluation of each module of the system according to a criterion adapted to the data given as input and to the results provided as output. Thus, according to the different modules, either the recall and precision measures were used, or the score defined by TREC.

We used recall and precision measures for both the question analysis and document selection modules. The question corpus was tagged by hand, and we judged the results with regard to this reference corpus. It should be noted that we did not evaluate named entity recognition within the documents. Indeed, for this purpose, we would have had to tag by hand the document corpus. We did it for the 500 questions of the TREC10 corpus, but we could not do it even for a large subset of the document corpus. Concerning the evaluation of the document selection, we used the reference data provided by the NIST. Indeed, the NIST provides the list of correct answers found by participants to TREC, and the documents where they have been retrieved, after each TREC conference it organizes. The NIST also provides the patterns of correct answers and a code which computes the score of the system. Concerning the sentence selection and answer extraction modules, we then used the score defined by TREC, so as to evaluate those modules in terms of the correct answers retrieved, either within the sentences, or in the final answers (less than 50 characters). In addition to the evaluation of each module, we performed a specific evaluation of the extraction patterns used in the answer extraction module, in order to assess the pattern relevance with regard to each question category.

3. Black box evaluation of the system

We present in table 2 the score obtained by the run that was sent to TREC10 evaluation. Strict and lenient scores correspond to the two judgements provided by human judges of the conference. These scores are slightly lower than the one computed with the code and data provided by the NIST. Indeed, computer matching of answer pattern cannot completely insure the correctness of the answer. It may happen that, although the correct words are retrieved, the document context shows that they do not constitute a correct answer.

Evaluation	strict	lenient	automatic
TREC10	0.181	0.192	0.205

Table 2: Evaluation of the overall system

The score of the system is rather weak. In order to better determine the weak points, we performed an evaluation of the overall system, for each question category.

The categories that we obtain from the question analysis module, are of very dissimilar size, from 2 questions for the smallest to 182 for the biggest. For more clearness, we brought together categories which had a number of shared caracteristics so as to create 12 sufficiently homogeneous categories. Nevertheless, we have to keep in mind that these categories include more specific sub-categories that we study with more detail in particular cases. Table 1 presents the score of the system attached to each question category, in decreasing order of the scores.

The *WhatbeNP* category has the biggest size. It includes a number of low size categories (such as *WhatbeNPforNP* for instance) that do not appear in the table. Only the *WhatbeNPofNP* appears because of its bigger size. Among the 492 questions from TREC10, only one is not taken into account in our statistics : *What causes gray hair?*, the only one instance in its category (*WhatverbNP*).

In this first approach, we note that the best scores are obtained by categories corresponding to a named entity. Either categories whose expected answer is, with very few exceptions, a named entity, (*Where*, a location, *When*, a date, *Who*, a person or an organization), or for which a great part of the questions expect a named entity as answer (*WhatNPdoNP*, *WhatbeNPofNP*, *How*). This result seems coherent in that the knowledge of the answer type allows a more precise location of the answer within the documents whose named entities are tagged.

4. Glass box evaluation

4.1. Question analysis evaluation

In our question answering system, the question analysis module is performed in order to assign the questions some features (see section 2.1.1.) that will be used in the answer module for the selection of candidate sentences and for the answer extraction. To find all these different items of information, we wrote rules using syntactic information of a shallow parser (Ait-Mokhtar and Chanod, 1997) and semantic information from the lexical base WordNet (Fellbaum, 1998). In the first time, we evaluated the performances of our question module for each feature using the precision measure, defined as follows:

A : the set of retrieved question features
Ra : the set of relevant question features within A
Precision : Ra / A

According to the table 3, our question module has rather good results. In this table, Category is the question category, NE-Type means the named entity answer type of the question, and Gen-type means the general semantic type of the question. Even if these results are rather good, it is necessary to evaluate more precisely this module in order to improve the recognition of question features which is essential in the following modules. But these results of the table 3 do not allow us to know what are the rules which must be refined or revised to improve the question module performances. So, to find them, we evaluated our question module for each question category. Indeed, the rules to find the different question features depend on the question's syntactic form, which corresponds to the question category. An evaluation of our question module according to the question category will allow us to detect if the rules which are written for these categories are sufficient and correct.

Category	NE-Type	Gen-type	Focus	Focus-Head
97.2%	90.5%	87%	85%	89.6%

Table 3: Evaluation of overall question module

This evaluation of our question module is performed by using the recall and precision measures for each question feature, defined as follows :

A : the set of retrieved question features
Ra : the set of relevant question features within A
R : the set of relevant question features
Precision : Ra / A
Recall : Ra / R

It is very interesting, with this evaluation, to observe that some categories, which obtained the good results in the overall system evaluation, have not necessarily good results in our question module. Indeed, according to results obtained for the question category *WhatNPdoNP* (see table 4), the recognition of question focus for this category is not good (59 % of precision and recall), even though in the overall system evaluation this category has a good result. These evaluations allow us so to detect that the rules for the question category *WhatNPdoNP* which recognize the question focus have to be refined or revised. According to overall system evaluation results for the question category *WhatNPdoNP* (see table 1), we would not have thought that their rules were wrong.

	Recall	Precision
Category	95.83 %	100 %
NE-type	93.33 %	93.33 %
Gen-type	100 %	100 %
Focus	59 %	59 %
Focus-Head	72.7 %	72.7 %

Table 4: Evaluation of question module for the category WhatNPdoNP

Moreover, if the category has a lower result in the overall system evaluation, this does not always imply that the results of question module evaluation will be also weak. For example, for the question category *WhatdoNP*, the precision and recall measures for the focus recognition (see table 5) are better than these of the *WhatNPdoNP* category, even though in the overall system evaluation (see table 1), this category has a wrong result. These results show us that the problem probably comes from the following modules.

	Recall	Precision
Category	95.4 %5	100 %
Focus	78.94 %	78.94 %
Focus-Head	84.21 %	84.21 %

Table 5: Evaluation of question module for the category WhatdoGN

On the opposite side, for the question category *Why*, we are certain that the recognition of question focus is a problem, as its recall and precision measures are weak (see table 6). These bad results could be caused by the rules which are weak or inexistent, but also by wrong results of the question syntactic analysis obtained by the shallow parser. Indeed, the shallow parser is not adapted to analyse questions. Sometimes, the parser does not recognize the verb of the question, or some noun phrases are incomplete, etc... For example, for the question *When did the Titanic sink?*, *sink* is recognized as the noun.

	Recall	Precision
Category	100 %	100 %
Focus	25 %	25 %
Focus-Head	50 %	50 %

Table 6: Evaluation of question module for the category Why

In the meantime, some question categories, for instance *WhatbeNPofNP* (see table 7), get good results in the recognition of different features. For the category *WhatbeNPofNP*, it is obvious that the modifications to improve the overall system evaluation results have to be performed in the subsequent modules.

	Recall	Precision
Category	100 %	96.15 %
NE-Type	100 %	100 %
Gen-Type	100 %	93.33 %
Focus	98 %	98 %
Focus-Head	96 %	96 %

Table 7: Evaluation of question module for the category WhatbeGNofGN

In conclusion, the evaluation per category allows us to know the weak points of our question module and more precisely the rules which have to be revised or refined or added to improve the black box evaluation.

4.2. Document selection evaluation

For the TREC10 evaluation campaign, the QALC system used the outputs provided by NIST, resulting from the application of their vectorial search engine on the document corpus for the set of questions. We evaluated this first selection by taking as relevance criterion the fact that a document actually includes the answer to the question. Evaluation is performed using the precision and recall measures, defined as follows:

R : the set of relevant documents
A : the set of retrieved documents
Ra : the set of relevant documents within A
Precision = Ra / A
Recall = Ra / R

Table 8 first shows the overall results, and then results about selected categories. Recall is in this case particularly low because the number of retrieved documents, 200 for each question, is very big compared to the total number of relevant documents. Indeed, the average number of relevant documents per question is only 8 out of the corpus of one million of documents. In the table 8, categories were put in the same order as in table 1.

Category	Precision	Recall
All categories	68 %	3 %
Where	59 %	5 %
WhatNPdoNP	63 %	3 %
WhatbeNPofNP	60 %	2 %
WhatbeNP	78 %	3 %
WhatdoNP	58 %	2 %
Why	89 %	1 %

Table 8: Evaluation of the first document selection

According to this table, categories which have the best score also have the best recall, but do not have the best precision. The score thus seems to depend more on recall than on precision. The density of relevant documents among the set of documents to process, is a factor which contributes to the success of answer seeking.

We then achieve the same evaluation for the selection which is subsequently performed by FASTR. Table 9 shows the results.

Category	Precision	Recall
All categories	47 %	6 %
Where	34 %	13 %
WhatNPdoNP	49 %	5 %
WhatbeNPofNP	50 %	6 %
WhatbeNP	54 %	7 %
WhatdoNP	45 %	5 %
Why	67 %	3 %

Table 9: Evaluation of the document selection through FASTR

This table confirms the results of the previous table. The final score is also better correlated with recall than with precision, though less than previously. The loss of precision that we observe is partly due to the weakness of the question tagging from the Treetagger. In particular, questions with the auxiliary *do* are ill-processed. For instance, in the question 950 *When did Elvis Presley die*, the word *die* is not tagged as verb but as noun. FASTR subsequently uses the tagging in order to build compound words from the question words, to assign a weight to each index within the documents, and to recognize the morphological and semantic variants of an index. In our example, compound words will thus be *Presley die* and *Elvis Presley die*, but not *Elvis Presley*. Nor shall we obtain the verb variants but the variants of the noun *die*. Finally, when the verb *die* will be found in a document, its index will get a weak weight because of its wrong category with regard to the category of the same word in the question. It should be noted that we recover this type of weighting error in the sentence selection module.

But, though there is a loss of precision with regard to the first document selection, previous evaluations of the QALC system show that the final score is higher when this second selection is performed (Berthelin et al., 2001). Indeed, the recall is higher after the FASTR selection, because of a decrease of the number of documents to process (about 60 documents per question instead of 200), and an improvement of the density of relevant documents among those ones.

4.3. Evaluation of answer processing

4.3.1. Comparative evaluation of the sentence selection and answer extraction modules

In order to evaluate the sentence selection module, we computed the score on the top 5 sentences that the module returns, as defined by TREC. Table 10 shows the results of the evaluation for each category. We also computed the number of correct answers retrieved per category, with no regard to their rank. The answer extraction is evaluated using these data: the evaluation measure is then the percentage of correct short answers extracted from the sentences which contain a correct answer.

Category	Sentence Score	Answer Score	Sentence -answer extraction
All categories	0.286	0.205	73%
WhatNPdoNP	0.425	0.272	53%
Where	0.415	0.316	93%
WhatNPvrbNP	0.333	0.167	50%
How	0.298	0.192	58%
WhatbeNPofNP	0.288	0.247	100%
Who	0.286	0.254	94%
WhatbeNP	0.281	0.189	69%
When	0.280	0.280	100%
WhatNPbeNP	0.274	0.182	63%
WhatdoNP	0.205	0.027	25%
Which	0.175	0.100	33%
Why	0.00	0.00	

Table 10: Evaluation of sentence selection and answer extraction processes

Items in table 10 are ranked according to the decreasing order of sentence scores. We then obtain a ranking rather different from the one which results from the decreasing order of final scores. Obviously, this is due to large differences in extraction ratio according to the categories.

Category *Why* is very small, only 4 questions that we do not answer even in terms of a sentence. As a matter of fact, this category has a bad recall though it has a good precision. It means that very few documents among the corpus contains an answer to these questions.

Results of table 10 show that the answer is correctly extracted when categories correspond to a named entity. Therefore, the good overall results of these categories are mostly due to successful extraction than to sentence selection correctness. Then, the *WhatbeNP* category has rather good results concerning the answer extraction process. Contrariwise, *WhatNPdoNP* and above all *WhatdoNP* category get lower extraction ratio. Questions from *WhatbeNP* and *WhatdoNP* categories, and partly from *WhatNPdoNP*, do not expect a named entity as answer. Thus, the answer extraction process uses extraction patterns in these cases. In order to determine the reason why performances are so different, we then have to study in more detail the answer extraction patterns used for these categories.

4.3.2. Answer extraction patterns evaluation

The evaluation of syntactic patterns used in the answer extraction process is performed on the ten most relevant sentences, for each question, resulting from the sentence selection module. During the answer extraction process, the QALC system note, for each sentence, the applied syntactic pattern and the type of focus used when applying this pattern (the focus itself, a proper name in the question or the general answer type). We then computed, by means of the correct answer patterns provided by the NIST, the number of correct answers with regard to the number of times the pattern was applied. The results of this evaluation are shown in table 11. In this table, items appears according to the decreasing order of pattern applying frequency. In the pattern expression, *Answer* means the noun phrase of the answer, *Focus* means the noun phrase of the focus, and the connecting element is indicated between them. The focus that appears in the pattern expression can be either the focus itself, or a proper noun of the question (*PN*), or the general semantic type of the expected answer (*gen*).

For instance, let us take the question 1008, *What is the Hawaii's state flower?*. The focus of the question, that has been determined by the rules of the question analysis module (see section 4.1.), is *Hawaii's state flower*, and the focus head is *flower*. The answer *Yellow hibiscus is the state flower of Hawaii* was extracted from the following candidate sentence: *Yellow hibiscus is the state flower of Hawaii, but Postrzech doesn't recommend them for evening luaus because they close at the end of the day* using the pattern *Answer be Focus*.

Category	pattern	focus	success
WhatbeGN	Answer*in*Focus	focus	6%
	Focus , Answer	focus	29%
	Answer , Focus	focus	6%
	Focus and Answer	focus	10%
	Focus be Answer	focus	22%
	A. such as Focus	focus	53%
	Answer (Focus	focus	50%
	Focus - Answer	focus	33%
	Focus (Answer	focus	100%
	Answer be Focus	PN	50%
WhatGNdoGN	Answer*in*Focus	gen	2%
	Answer*in*Focus	focus	0%
	Answer*in*Focus	PN	8%
	Answer , Focus	focus	16%
WhatdoGN	Answer*in*Focus	focus	0%
	Focus , Answer	focus	33%

Table 11: Success ratio of answer extraction patterns

Two main results appears when studying this table: on the one hand, the most frequently recognized pattern, *AnswerinFocus*, is the one which has the lower success ratio, and on the other hand, the most precise patterns, in particular those which have no very frequent connecting element (parenthesis or *such as*), produce the most correct answers. In fact, during the pattern recognition process, patterns are matched with sentences in a pre-defined order, and the first recognized pattern is kept. The *AnswerinFocus* pattern is most of the time the first to be matched. As it frequently appears in documents, it ends up being also the most frequently recognized. We performed a new run after having changed the pattern matching order, putting the *AnswerinFocus* pattern at the last position. We thus retrieved new correct answers.

For instance, the question 1265, *What currency do they use in Brazil?*, did not get the correct answer *the Real* from the following sentence: *During the interview Mr Ricupero suggested he, and the government, were using Brazil's latest anti-inflation plan and its main component , a new cur-*

rency , the Real, to help Mr Cardoso win votes .

The pattern *AnswerinFocus* gave as answer: *a new currency*. With the new pattern matching order, the correct answer *a new currency , the Real* was matched with the *Focus , Answer* pattern and the general answer type *currency* as focus. The black box evaluation of the run performed on the TREC10 corpus with the new pattern matching order gives a score that is 3% above the previous score. This is a small improvement, but it therefore emphasizes the importance of one factor, the pattern matching order.

Obviously, there are other most important reasons which account for the difference between the categories. First, very few patterns are recognized concerning the *WhatdoGN* and *WhatGNdoGN* categories. We certainly have to introduce new patterns. Nevertheless, we found out that some types of answer could not be reduced to a pattern, but could have been retrieved by means of a complete syntactic analysis which also produces the syntactic dependencies. For example, the answer to the question 282, *What do ladybugs eat?* (category *WhatdoNP*), which is *the aphids*, is not easy to extract from *Bailey recommended turning ladybugs loose in the garden to eat the aphids that may appear*. We have to detect the syntatic dependency between *ladybugs* and *eat*.

5. Results analysis

The glass box evaluation of the QALC system shows that the results of each module have an effect upon the following modules. In fact, the data produced by one module are actually used in other modules, and as a result, errors are propagated. Black box and glass box evaluations according to the question category both show that the system behaviour depends on the category. The black box evaluation shows the differences, while the glass box evaluation shows that the differences vary along with the modules. We now see in more detail those two issues.

5.1. Relationships between the modules

We saw that the results from the question analysis are used in two other modules, on the one hand in the document selection, based on the term extraction, and on the other hand in the answer extraction, based on question category and focus. The errors that are produced in the analysis of the question are hardly recovered in the following modules. For instance, if the analysis does not find the focus of the question, and if there is no proper name in the question, neither an expected general type of answer, then no pattern could be applied.

Nevertheless, we note some stability in the performance of the different modules with regard to the different corpora. For instance, the answer extraction module gets a success ratio, over all the categories, which is rather stable. Indeed, it gets 73% of success on the TREC10 corpus (see section 4.3.1.). On the first 200 questions of the TREC9 corpus, it gets 76% of success, and finally on the TREC9 corpus, when only the sentences including correct answers are taken as document corpus, the success ratio is of 78%. It should be noted that the success ratio is slightly higher on the TREC9 corpus, but the syntactic patterns were built from this corpus. Thus, if the performance of the sentence selection increases, then the overall score will also increase.

5.2. Question categorization

We saw in section 4.3.2., that the most precise patterns obtained the best results. Therefore it seems important that the question categorization should be as precise as possible. At present, the categorization is mainly based on the syntactic form of the question, and can be refined. For instance, the syntactic form WhatbeNP correspond to two different refined forms. The first one, *WhatbeNP*, for example *What is epilepsy?*, corresponds to a request for a definition. The second one, *WhatbeDefiniteNP*, for example *What is the brightest star?*, corresponds to a request for an instance of a definite object. Both cases do not correspond to the same patterns. For instance, the patterns which have as connecting element *or* or *such as* can be applied in the first case, but not in the second one. Therefore, we will have to split these two categories in order to obtain better results.

6. Conclusion

In this paper, we presented an evaluation of a question answering system, based on an approach which combines a black box and a glass box evaluation. In addition, we evaluate the system according to a corpus categorization. The various methods that we set allow us to control the changes and improvements in our system, as shown by the example on the pattern order in section 4.3.2..

When we consider the results on the different modules over all categories, we can see that the sentence selection module obtains the lowest results. We, and the team who develops the question answering project in our laboratory, are currently working on the improvements that have to be done concerning this module. The different results of our evaluations also show research directions that may lead to the improvement of the other modules: modifications in the rules of the question module in order to categorize the questions more precisely, creation of new syntactic patterns and improvement of their ranking, and finally, use of the syntactic dependencies in the question.

7. References

S. Ait-Mokhtar and J. Chanod. 1997. Incremental finite state parsing. *In Proceedings of ANLP-97*.

J-B. Berthelin, B. Grau, and M. Hurault-Plantet. 2001. Two levels of evaluation in a complex nl system. *In Proceedings of ACL2001 Worshop on Evaluation Methodologies for Language and Dialogue Systems*.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press , Cambridge, ma edition.

O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, and C. Jacquemin. 2001a. Document selection refinement based on linguistic features for qalc, a question answering system. *In Proceedings of RANLP2001*.

O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, and A. Vilnat. 2001b. Finding an answer based on the recognition of the question focus. *In Proceedings of TREC 10*.

C. Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. *In Proceedings of ACL'99*, 1:341–348.

A. Simpson and N. Fraser. 1993. Black box and glass box evaluation of the sundial system. *In Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 1423–6.

E. Voorhees and D. Tice. 2000. Implementing a question answering evaluation. *In Proceedings of the second International Conference on Language Resources and Evaluation (LREC)*.