Prodicos experiment feedback for QA@CLEF2006

E. Desmontils, C. Jacquin & L. Monceaux
LINA, University of Nantes
2, rue de la Houssinière, BP92208
F-44322 Nantes CEDEX 3, France
{Emmanuel.Desmontils,Christine.Jacquin,Laura.Monceaux}@univ-nantes.fr

Abstract

We present the second version of the Prodicos query answering system which was developed by the TALN team from the LINA institute. We have participated to the monolingual evaluation task dedicated to the French language. We firstly present the question analysis step which makes it possible to extract many features from the questions (question category, question type, question focus, answer type, ...). For this new campaign, new features are extracted from the questions in order to improve the passage selection process (named entities, noun phrases and dates). We also determine four different strategies that will be used during the answer extraction step (entity named strategy, numerical entity strategy, acronym definition strategy, pattern-based strategy). We also take into account a new category of question (lists). We then present the passage selection process whose goal is to extract from the journalistic corpora the most relevant passages which answer to the question. This year, we present a new strategy applied to definitional queries. We use external knowledge (Wikipedia encyclopedia) to add information to these kinds of questions. Then, we discuss, in details, the major improvements made on our system at the answer extraction module level. According to the strategies determined during the question analysis stage, we present the 4 different strategies applied to this step. We present, in details and independently, each strategy and their use context. Afterwards, for the passage selection and answer extraction modules, the evaluation is put forward to justify the results obtained.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Question analysis, Passage selection, Answer extraction, Pattern-based answer extraction

1 Introduction

In this paper, we present the second version of the Prodicos query answering system which was developed by the TALN team from the LINA institute. It was our second participation to the QA@CLEF evaluation campaign. We have decided to participate to the monolingual evaluation

task dedicated to the French language. This campaign enables us to analyse the performances of our system. Firstly, we present the various modules constituting our system and for two of them (passage extraction module and answer extraction module) , the evaluation is put forward to justify the results obtained.

2 Overview of the system architecture

The Prodicos query answering system is divided into three parts (figure 1):

- question analysis module;
- passage extraction module (extracts passages which might contain the answer);
- answer extraction module (extracts the answer according to the results provided by the previous module).

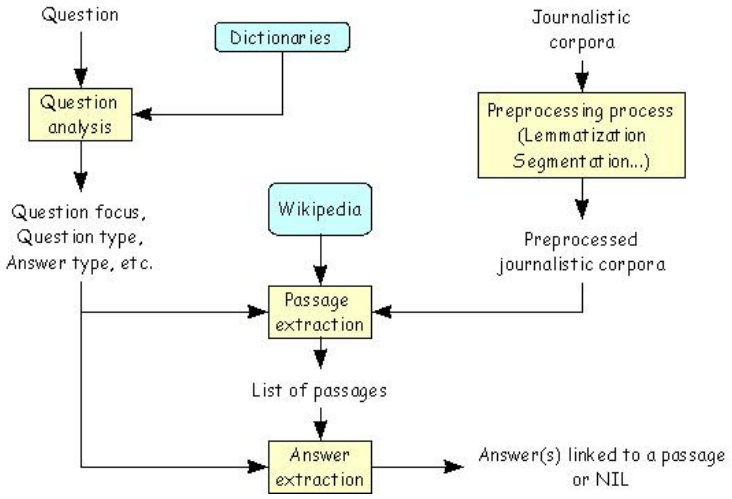

Figure 1: The global architecture

The modules of the Prodicos system are based on the use of linguistic knowledge, in particular lexical knowledge coming from the EuroWordnet thesaurus [6] and syntactic knowledge coming from a syntactic chunker which has been developed by our team (by the use of the TreeTagger tool [5]).

The system has participated to the QA@CLEF 2006 evaluation campaign for the monolingual query answering task dedicated to the French language. This campaign enables us to make an evaluation of the system. We present, in the next sections, in greater detail, the various modules which belong to the Prodicos system and the linguistic tools used to implement them. In parallel, we analyse in detail the results for the passage extraction module and the answer extraction module. The question analysis module was been evaluated last year during the QA@CLEF 2005 campaign [1].

3 Question analysis module

The question analysis module aims to extract relevant features from questions that will make it possible to guide the passage selection and the answer search. We extract many features from the questions [1]: question category, question type, question focus, answer type, principal verb, etc.

The question category is determined according to specific syntactic rules. The main feature which comes from the question analysis is then the question type. It will not only help to determine the strategy to perform an answer search but also it will make it possible to select rules to extract other important features from questions (answer type, question focus). We defined twenty question types which correspond to a simplified syntactic form of the question [1] (for example the type `QuiVerbeGN`). The question type makes also it possible to verify the answer type that will be retrieved. The answer type may be a named entity (Person, Location-State, Location-City, Organization...), or a numerical entity (Date, Length, Weight, Financial-Amount...). The question focus corresponds to a word or a word group involved in the question. Its main particularity is that, generally around it, the answer is present within the passages which may contain the answer. These different features are extracted by using the TreeTagger tool and then, according to the part-of-speech tags, by building some rules to determine the question chunks (noun phrase, adjective phrase, adverb phrase, prepositional phrase, verb phrase). Then, according to the previous syntactic chunks, we have written rules which make it possible to extract, from the questions, information like question focus, principal verb,... For determining answer type, we use semantic knowledge (EuroWordnet Thesaurus). We build lists of words which are hyponyms of some predefined words which are considered like categories and we use them in order to generate the answer type [3], [1].

For this new campaign new features are extracted from the questions in order to improve the passage selection process (named entities, noun phrases and dates). For example, for the queries: "Qui est Boris Becker ?" (the 90th question: "who is Boris Becker?") and "Qu'est-ce que l'effet de serre ?" (189th question: "what is the greenhouse effect?"), we now consider "Boris Becker" and "effet de serre" as a single entity. We also determine a new feature which is the strategy to use to search the right answer. It is determined according to the question focus and the question type. These strategies are either an entity named strategy, either a numerical entity strategy, either an acronym definition strategy or a pattern-based strategy (Figure 2). For example, if we take into account the first case, this means that the answer extraction module must use a named entity recognizer in order to extract the answer ...

```
<QUESTION NumQuest="0007">
 <ANALYSE>
  <TEXTE_QUESTION lang="F">Quel pays l' Irak a -t-elle envahit en 1990 ?</TEXTE_QUESTION> ...
  <CATEGORIE>QuelGN</CATEGORIE>
  <STRATEGIE>Entité Nommée</STRATEGIE>
  <DATE>1990</DATE>
  <NBRE_REPONSE>1</NBRE_REPONSE>
  <TYPE_EN><EN>LOCATION-STATE</EN></TYPE_EN>
  <LISTE_NP><NP>Irak</NP></LISTE_NP>
  <FOCUS lang="F" trad="N">
    <LEMMES><LEMME>pays</LEMME></LEMMES>
    <TETE forme="pays" eti="NN" leNum="F2"><LEMME>pays</LEMME></TETE>
  </FOCUS>
  </ANALYSE>
</QUESTION>
```

Figure 2: Example of a question analysis

For questions whose answer is of type list, we generate a new feature which is the number of answer which we are waiting for. In this context, we have three kinds of question, classified according to the number of awaited answers. Answer of:

- one answer type. For example "Qui est Boris Becker ?" ("Who is Boris Becker?"),
- precise number answer type (often extracted from the noun phrase corresponding to the question focus). For example, for the question: "Qui sont les deux principaux responsables

[1] excepted for definitional questions [3]

de l'attentat d'Oklahoma City ? (the 92 th question: "who are the two persons in charge for the terrorist attack of Oklahoma City?"), the question focus is "les deux principaux responsables de l'attentat d'Oklahoma City" ("two persons in charge for the terrorist attack of Oklahoma City"),

- several answers type (undefined number,often extracted from the noun phrase corresponding to the question focus). For example, for the question: "Citer le nom de tous les aéroports de Londres, en Angleterre." (the 88th question: "give the name of all London's airport, in England), the question focus is "le nom de tous les aéroports de Londres, en Angleterre" ("the name of all London's airport").

4 Passage selection module

The goal of this module is to extract from the journalistic corpora the most relevant passages which answer to the question (ie, the passages which might contain the answer). Firstly, the corpora are processed and marked with XML annotation in order to locate the passages. The corpora are then annotated with part-of-speech and lemma by using the TreeTagger tool. A passage is often a sentence excepted for example for a person's citation which is a set of sentences whose union is regarded as a single passage.

Then, the corpora are indexed by the Lucene search engine[2]. The indexing unit used is the passage. For each question, we then build a Lucene request according to the data generated by the question analysis step. The request is built according to a combination of some elements linked with the "or" boolean operator. The elements are: question focus, named entities, principal verbs, common nouns, adjectives, dates and other numerical entities. For a particular request, the passage extraction module provides a sorted passage list which answers to the request. The sort criterion is a confidence coefficient associated with each passage in the list. It is determined according to the number and the category of the question elements which are found in passages. For example, if the question focus belongs to a passage, the confidence coefficient of this passage is high, because the question focus is very important for the answer extraction step [1]. When the passage extraction module stops, only the 50 passages with the highest confidence coefficient are kept.

We have a particular strategy for the definitional questions. We use external knowledge to add information to these kinds of questions. The external source used is the Wikipedia encyclopedia. We expand the question focus (which is either a simple noun or a complex noun phrase) with pieces of information extracted from Wikipedia articles. In this aim, we add to lucene request the noun phrases or simple nouns which belong to the first sentence of the corresponding Wikipedia article (if the question focus exists in the encyclopedia or if the noun phrase whose we search the definition is not polysemous). The added noun phrases are determined with the help of the TreeTagger tool. For example for the 4th question "Qui est Radovan Karadzic ? " (Who is Radovan Karadzic?) , the query sent to lucene search engine is : "Radovan" OR "Karadzic" OR "Radovan Karadzic" OR "homme politique" ("politician") OR "psychiatre" ("psychiatrist").

After the CLEF 2006 evaluation campaign, we have made the study, for the definitional queries, of the position of the first passage belonging to the list of returned passage, which contains the right answer (table 1).

The set of Clef evaluation queries comprises 40 definitional queries. For only eight of them, the question focus does not belong to the Wikipedia encyclopedia. This shows that the queries are often general and in this context, the French version of the encyclopedia has a good coverage. Three of them belong to the "wrong tagged" category. For these questions, the TreeTagger tool gives a wrong part-of-speech for the question focus and the consequence is, that the passage extraction process does not correctly detect the right passage for the answer. These question focuses are: "Euro Disney", "Crédit Suisse" and "Javier Clemente". These named entities are each divided into two single words and some of these words have a wrong part-of-speech associated to

[2]http://lucene.apache.org/java/docs/

	First passage	**coef=1 & not first passage**	**coef < 1**	**Not present in any passage**	**Total**
Not in Wikipedia	3	3	2	0	8
Wrong tagged	0	1	1	1	3
Not present in corpus	0	0	0	2	2
Other cases	15	5	3	4	27
Total	18	9	6	8	40

Table 1: Passage extraction evaluation for definitionnal queries

them. Indeed, for French language, some words constituting these named entities are ambiguous and represent either an adjective or a common noun. In table 1 , we can see that for 67% of the definitional questions, the passages containing the right answer take the value 1. Moreover, for 83% of them the answer belongs to a passage selected during this step. This are good results and this shows that the use of encyclopedic knowledge helps the selection passage process. The nature of the resource (Wikipedia) is also very interesting because of the recurrent problem for French language to have such kind of resource at one's disposal. The multilingual property can also be used in a cross-language evaluation context.

5 Answer extraction

5.1 Global process

This step comes at the end of our process. After the question analysis and the passage selection, we have to extract correct answers corresponding to questions. To this end, we use on the one hand elements coming from the question analysis like, for instance, the question's category, the strategy to use it, the number of answers, and so on (see figure 2 for an example of a part of such an analysis, element shown are used in this step) and, on the other hand, a list of passages selected and evaluated by our previous step according to this question.

The goal of this step is to find the precise answer(s) to a question. An answer is built with the answer itself, the passage used to answer and, a trust value. This ending process can be divided into 4 local steps (figure 3):

1. according to the question's strategy, the convenient entity extraction module is selected,
2. candidate answers are detected and selected by the previous selected module,
3. answers are evaluated and the answer(s) with the highest trust coefficient is (are) kept,
4. passages where each answer has been found are also associated to the selected answer.

The question's analysis can give 4 groups of categories which correspond to 4 possible strategies: numerical entities extraction, named entities extraction, acronym definitions extraction and pattern-based extraction (the default one). Now, we will present processes associated to each strategy and the build of final answers.

5.2 Numerical entities extraction

For locating numerical entities, we use a set of dedicated regular expressions. These expressions make it possible to the system to extract numerical information namely: dates, duration, times, periods, ages, financial amounts, lengths, weights, numbers and ratios. It uses the MUC (Message Understanding Conference) categories ("TIMEX" and "NUMEX") to annotate texts. For example, lets take the 13th question: «En quelle année la catastrophe de Tchernobyl a -t-elle eu lieu ?» (the year of the Tchernobyl's nuclear explosion). Our numerical extraction tool gives results as for the 7th sentence of "LEMONDE95-041936" shown in figure 4.

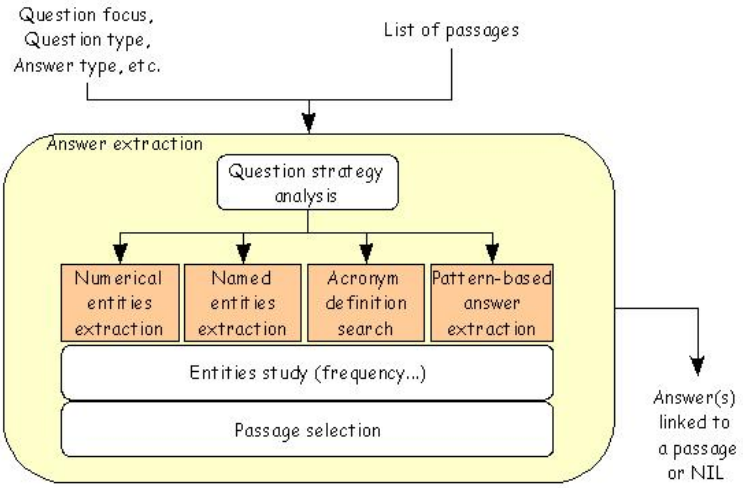

Figure 3: The answer extraction process

```
<enonce num="7" num_doc="LEMONDE95-041936" coef="0.90909094">
 <texte>
 De l' accident de Three-Mile Island aux Etats-Unis en 1979 en passant par la catastrophe de Tchernobyl
 en Ukraine en 1986 , est apparue la nécessité d' un renforcement de la sécurité et d' une standardisation
 des choix de sûreté des centrales .</texte>
 <texte-EN>
   De l' accident de Three-Mile Island aux Etats-Unis <timex type="date">en 1979</timex>
   en passant par la catastrophe de Tchernobyl en Ukraine <timex type="date">en 1986</timex> , est apparue
   la nécessité d' un renforcement de la sécurité et d' une standardisation des choix de sûreté  des centrales .
 </texte-EN>
</enonce>
```

Figure 4: Example of a numerical entities search

5.3 Named entities extraction

For locating named entities, NEMESIS tool [2] is used. It was developed by our research team. Nemesis is a French proper name recognizer for large-scale information extraction, whose specifications have been elaborated through corpus investigation both in terms of referential categories and graphical structures. The graphical criteria are used to identify proper names and the referential classification to categorize them. The system is a classical one: it is rule-based and uses specialized lexicons without any linguistic preprocessing. Its originality consists on a modular architecture which includes a learning process. For example, lets take the 7th question: «Quel pays l' Irak a -t-il envahi en 1990 ?» ("Which country Iraq did it invade in 1990?"). Figure 5 show what NEMESIS gives as results for the 21th sentence of "LEMONDE95-040819". It detects two country names ("Irak" and "Koweït") and a people's proper name ("Yasser Arafat").

5.4 Acronym definition extraction

For acronym's definition search, we use a tool developed by E. Morin [4] based on regular expressions. It detects acronyms and links them to their definition (if it exists). For example, lets take the 28th question: «Qu'est-ce que l' OMS ?» ("What means OMS?"). This tool gives results as the one of the figure 6 that shows the analysis the 20th sentence of "ATS.941027.0143"[3].

[3]In this example "SIGLE" describes an acronym and "DEF" (with the same identifier "no") its definition

```
<enonce num="21" num_doc="LEMONDE95-040819" coef="0.85714287">
 <texte>1990 . 2 août : invasion du Koweït par l' Irak , soutenu par Yasser Arafat .</texte>
 <texte-EN>
  1990 . 2 août : invasion du  <NP Categorie="Pays" Classe="Toponyme" Id="1">Koweït</NP>
  par l' <NP Categorie="Pays" Classe="Toponyme" Id="2">Irak</NP> , soutenu par
  <NP Categorie="Patronyme" Classe="Anthroponyme" Id="3">Yasser Arafat</NP> .
 </texte-EN>
</enonce>
```

Figure 5: Example of a named entities search

```
<enonce num="20" num_doc="ATS.941027.0143" coef="1.0">
 <texte>
 Une visite touristique sur ces places ne présente cependant aucun danger , de même qu' un repas à base
 de poisson , à condition qu' il soit cuit ou frit , selon une responsable de l' Organisation mondiale de la santé ( OMS ) .
 </texte>
 <texte-EN>
  Une visite touristique sur ces places ne présente cependant aucun danger , de même qu' un repas à base
  de poisson , à condition qu' il soit cuit ou frit , selon une responsable de l'
  <DEF no="1" sigle="OMS">Organisation mondiale de la santé</DEF> ( <SIGLE no="1">OMS</SIGLE> ) .
 </texte-EN>
</enonce>
```

Figure 6: Example of an acronym definition search

5.5 Pattern-based answer extraction

For the pattern-based answer extraction process, we developed our own tool. According to question categories, syntactic patterns were defined in order to extract answer(s) (see figure 7 for an example of a pattern set associated with the question category called "Definition"). These patterns are based on the question focus and makes it possible to the system to extract the answer. Patterns are sorted according to their priority, ie answers extracted by a pattern with an higher priority are considered as better answers than the ones extracted by patterns with a lower priority.

```
    <regle>
      <Categorie> Definition </Categorie>
      <patron> GNRep GNFocus </patron>
      <patron> GPRep GNFocus </patron>
      <patron> GNFocus GNRep </patron>
    </regle>
```

Figure 7: Example of patterns associated to a question category

As a result, for a given question, patterns associated with the question category are applied to all selected passages. Thus, we obtain a set of candidate answers for this question. For example, lets take the 2nd question: «Qu'est ce que Hubble ?» ("What is Hubble?"). This question corresponds to the category "Definition" (see figure 7). Pattern-based extraction process gives a set of candidate answers as the one presented in figure 8.

Patterns (syntactic patterns) are based on the noun phrase that contains the focus of the question. Therefore, the first step consists in selecting only passages which could contain the answer and which contain the focus of the question. To apply syntactic patterns, passages are parsed and divided into basic phrases such as noun phrase (GN), adjectival phrase (GA), adverbial phrase (GR), verb phrase (NV), etc. We use a parser which is based on TreeTagger tool for annotating text with part-of-speech and lemma information. Subsequently, passages are studied to detect the focus noun phrase and to apply each pattern of the question's category. The figure 9

```
<QUESTION num="0002" nbrep="1" categorie="Definition">
  <REP num="1" doc="LEMONDE95-040794-1.0"> l' objectif </REP>
  <REP num="1" doc="LEMONDE95-023629-1.0"> le télescope </REP>
  <REP num="1" doc="LEMONDE95-033842-1.0"> par le télescope </REP>
  <REP num="1" doc="ATS.940217.0089-1.0">la réparation</REP>
  <REP num="1" doc="LEMONDE95-023628-1.0"> le télescope  sa position</REP>...
</QUESTION>
```

Figure 8: Example of noun phrases extracted using our pattern-based extraction algorithm

```
... <Groupe type="PV" id="E-3G8">
  <F id="E-3F0">à</F> <F id="E-3F1">équiper</F>
</Groupe>
<Groupe type="GR" id="E-3G9">
  <F id="E-3F0">aussi</F>
</Groupe>
<Groupe type="GN" id="E-3G10">
  <F id="E-3F0">le</F> <F id="E-3F1">télescope</F>
</Groupe>
<Groupe type="GP" id="E-3G11">
  <F id="E-3F0">de</F> <F id="E-3F1">Hubble</F>
</Groupe>
<Groupe type="GP" id="E-3G12">
  <F id="E-3F0">d'</F> <F id="E-3F1">une</F> <F id="E-3F2">optique</F>
</Groupe> ...
```

Figure 9: The first syntactically annotated sentence containing the answer of the question 2.

gives an example of an annotated sentence for the question 2 («Qu'est ce que Hubble ?»). In this case, all patterns, for the category "Definition" (see figure 7), are applied. The pattern "GNRep GNFocus" can be applied (the answer focus is "Hubble"). Thus, the noun phrase "E-3G10" is a candidate answer.

For the "Definition" category, the pattern strategy gives good results. Nevertheless, for more complex questions, a semantic process could improve the answer search. Indeed, sometimes the build of powerful patterns is a difficult task (such as for the question «Dans quel lieu des massacres de Musulmans ont-ils été commis en 1995 ?») knowing our patterns are based on the question's focus and the focus is not always easy to find. In addition, the answer type is not always easy to define without semantic information. Another improvement can take into account verb categorization. Indeed, the verb in the question is quite important.

5.6 Answer selection

When the answer type was been determined by the question analysis step, the process extracts, from the list of passages provided by the previous step, the candidate answers. Named entities, acronym definitions or numerical entities closest to the question focus (if this last is detected) are supported. Indeed, in such cases, the answer is often situated close to the question focus.

The answer selection process depends on the question category. For numerical entities, named entities and acronym definitions, the right answer is the one with the best frequency. This frequency is weighted according to several heuristics such as: the distance (in words) between this answer and the question focus, the presence in the sentence of named entities or dates from the question, etc. For answers extracted by the pattern-based selection, two strategies are used according to the question category:

- the selection of the first selected answer obtained by the first applicable pattern,
- the selection of the most frequent answer (the candidate answer frequency).

Most of the time, the first heuristic is the better one. Indeed, the selected answer is the first one obtained by the first applicable pattern (patterns sorted according to their convenience) and into the first passage (sorted by the passage selection step according to their convenience). Nevertheless, for definitionnal questions such as the 90th question «Qui est Boris Becker ?» ("Who is Boris Becker?") or the first question «Qu'est ce qu'Atlantis ?» ("What is Atlantis?"), we noted that the better strategy is the candidate phrase frequency. Indeed, for this question category where the number of question's terms is low, the passage selection step does not make it possible to the system to select with precision passages containing the answer. Therefore, the frequency-based strategy generally selects the right answer. For example, for the second question, the answer «télescope» is selected for the definitionnal question (figure 9) because of its frequency. Table 2 presents all results of our run according to question types.

Question type	**R**	**U**	**X**	**W**	**Total**
Named entities extraction	23	2	3	53	81
Numerical entities extraction	15	0	4	24	43
Acronym definition search	4	0	0	1	5
Pattern-based answer extraction	16	0	11	44	71
Total	58	2	18	122	200

Table 2: Results synthesis

43 questions (21.5%) was considered as named entities extraction strategy. Our process find at least 15 right answers. 81 questions (40.5%) was considered as named entities extraction strategy. Our process find at least 23 right answers. 71 questions (35.5%) was considered as pattern-based answer search type. Our process finds at least 16 right answers.

5 questions (2.5%) were analyzed as acronym definition search: question 28 (OMS), question 48 (OUA), question 95 (RKA), question 129 (KMT) and question 145 (TDRS). Our system found 4 good answers (28 with "Organisation Mondiale pour la Santé", 48 with "Organisation de l' unité africaine", 129 with "Kouomintang" and 145 with "Tracking and Data Relay Satellite"). Only "RKA" (for "Agence Spatiale Russe") was not found. This is a specific case. In fact, the definition does not contain the letter "K". Actually, "RKA" is based on the russian definition that does not appear in the corpus.

For 4 questions (5 questions awaiting a list were undetected), a list of answers is awaited:

- question 88 (Pattern-based search strategy), «Citer le nom de tous les aéroports de Londres , en Angleterre .» which demands an unlimited list,
- question 92 (Named entities extraction strategy), «Qui sont les deux principaux responsables de l' attentat d' Oklahoma City ?», waiting for 2 answers,
- question 100 (Named entities extraction strategy), «Donner le nom des neuf planètes qui constituent le système solaire .», 9 answers awaited.
- question 117 (Named entities extraction strategy), «Quels sont les sept pays les plus industrialisés du monde ?», 7 answers awaited.

In such cases, our process does not produce satisfactory results. For the first "question" (88), a process problem due to the unlimited list causes no answer ! For the second one (92), we give the right answers (with the same associated sentence). For the third one (100), we give only one wrong answer. For the last one (117), we have found 3 of the seven answers, ie (Canada, Russia, Bosnia, Italy, USA, Ukraine, Uruguay) instead of (USA, Canada, Japan, United Kingdom, France, Germany and Italy).

6 Conclusion

In this experiment report, we have studied our second version of the Prodicos QA system on QA@CLEF2006 question set. The comparison between this version and the first one studied on QA@CLEF2005 is not an easy task. Indeed, question types are quite different. For instance, in the CLEF'2005's session, 21 questions were acronym definition search. Conversely, in the CLEF'2006's session, only 5 questions were acronym definition search. Consequently, we have not presented our result relatively to the preceding system.

The hard result of the evaluation of the 2006 session is a rate of good answers (overall accuracy) of 29% (14.5% at the 2005 session). We regard this result as encouraging (although definitely perfectible). In addition, acronym definition questions are less numerous while our tool is more powerful. The rate of good answers is definitely low but some improvements were made. For instance, the pattern-based answer extraction found 16 answers whereas it found only 2 answers last year. In addition, in this process, 11 answers are "inexacts" (int the set "X"). As a result, it is obvious that the syntactic parser has to be improved to reduce the set "X" and, consequently, to increase the set "R".

Positive points are: (1) a good study of the question type, (2) a correct passage search and (3) an improvement of the answer extraction process. Nevertheless, some improvements have to be done concerning: (1) the question focus identification, (2) the use of semantic resources in French language for all process steps and (3) the answers extraction processes. Furthermore, we have to improve our French semantic ressource. Indeed, EuroWordnet in its French version has some defaults like the lack of definitions for concepts, relations between some concepts are unavailable, etc.

References

[1] L. Monceaux, C. Jacquin, and E. Desmontils, "The query answering system Prodicos", C. Peters, F. C. Gey, J. Gonzalo, G. J.F. Jones, M. Kluck, B. Magnini, H. Müllern and M. de Rijke eds, Proceedings of Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Revised Selected Papers, Vienna, Austria, September 2005, volume 4022, LNCS, Springer Verlag, forthcoming.

[2] N. Fourour, "Identification et catégorisation automatiques des entités nommées dans les textes français", These en informatique, Université de Nantes, LINA (2004)

[3] L. Monceaux, "Adaptation du niveau d'analyse des interventions dans un dialogue - application à un système de question - réponse", These en informatique, Paris Sud, ORSAY, LIMSI (2003)

[4] E. Morin, "Extraction de liens sémantiques entre termes à partir de corpus de textes techniques", Thèse en Informatique, Université de Nantes, LINA, Décembre 1999. http://www.sciences.univ-nantes.fr/info/perso/permanents/morin/article/morin-these99.pdf

[5] H. Schmid, "Improvements in Part-of-Speech Tagging with an Application To German". In S. Armstrong, K. W. Chuch, P. Isabelle, E. Tzoukermann & D. Yarowski (Eds.), Natural Language Processing Using Very Large Corpora, Dordrecht, Kluwer Academic Publisher, 1999

[6] P. Vossen "EuroWordNet: A Multilingual Database with Lexical Semantic", editor Networks Piek Vossen, university of Amsterdam, 1998.