The Answer Validation System ProdicosAV

Christine Jacquin, Laura Monceaux, Emmanuel Desmontils
Université de Nantes , Laboratoire LINA ,2 rue de la Houssinière, BP92208, 44322 Nantes cedex 03, FRANCE
christine.jacquin,laura.monceaux,emmanuel.desmontils@univ-nantes.fr

Abstract

In this paper, we present the ProdicosAV answer validation system which was developed by the TALN team from the LINA institute. The system is composed of a question analysis module, a ranking passage module, an answer extraction module and an answer validation module.

Keywords

Question answering, Temporal Validation, Answer Validation

1 Introduction

In this paper, we present the ProdicosAV system which was developed by the TALN team from the LINA institute and which participated to the Answer Validation Exercice for French. This system is based on the PRODICOS System which participated two years ago to the Question Answering CLEF evaluation campaign for French. The ProdicosAV system is composed of four modules whose some of them come from the PRODICOS system. We present the adaptation of these four modules for the AVE 2008 campaign and the ProdicosAV results for the AVE 2008 campaign.

2 Overview of the system architecture

The ProdicosAV system is divided into four parts :

- question analysis module;
- passage ranking module (ranks passages according to their ability to contain the answer);
- answer extraction module (extracts the candidate answers from passages and ranks them according to the results provided by the previous module).
- validation module (compares the answers given by the previous module with those proposed by the test data and takes the decision if the answer is selected or validated or rejected)

We present, in the next sections, the various modules which belong to the ProdicosAV system.

3 Question analysis module

The question analysis module aims to extract relevant features from questions that will make it possible to guide the passage ranking and the answer search. We extract many features from the questions [7]:

- question type

- question focus
- answer type
- strategy

The first and main feature which comes from the question analysis is the question type. Twenty question types are defined which correspond to a simplified syntactic form of the question [1] (for example the type `QuiVerbeGN`). The question type 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). The answer type may be a named entity (Person, Location-State, Location-City, Organization...), or a numerical entity (Date, Length, Weight, Financial-Amount...). For determining answer type, semantic knowledge coming from EuroWordnet Thesaurus [1] is used. Lists of words are build which are hyponyms of some predefined words which are considered like categories and are used in order to generate the answer type [4]. The question focus corresponds to a word or a group of word involved into the question. It is generally located close to the answer within the passages which may contain the answer. The strategy is a criteria use to search the right answer. It is determined according to the question focus and the question type. The strategies available are either an named entity strategy, either a numerical entity strategy, either an acronym definition strategy or a pattern-based strategy. Other features are extracted from the questions in order to improve the passage ranking process (named entities, noun phrases and dates). For example, for the query "Quand Abagelard de Paris est-il né ?" , "Abagelard de Paris" is considered as a single entity.

For example, if the question is "*Quand Abagelard de Paris est-il né ?*", the analysis of this question is:

1. Question type: `QUAND`
2. Answer type: `DATE`
3. Question Focus: *Abagelard de Paris*
4. Strategy: *Numerical Entity*
5. Named Entity: "Abagelard de Paris"

4 Passage ranking module

The role of this module is to rank the passage acording to their ability to contain the answer to the question. The strategy used relies on a density measure. [6] made a quantitative evaluation of passage retrieval algorithms for question answering and they showed that systems based on density measure perform better than the ones based on other techniques. The density measure approach lies on a scoring function based on how close keywords appear to each other. In our evaluation, we only use a density measure to rank the given passage according to selected words from the query but not to extract passage from large document.

For each question a kind of request is built according to the data generated by the question analysis step. The request is composed of a combination of elements such as question focus, named entities, principal verbs, common nouns, adjectives, dates and other numerical entities. These elements are also weighted according to their importance for determining the potential answer. The weight of each element depends on the question types. For example for a question type equals to "date", the coefficient associates with a date element is greater than the one linked with a principal verb element coefficient. The density measure used is based on the one from [5] but some adjustment were made.

For all query passage, let m be the number of query terms belonging to the passage and let k be the number of words belonging to the passage. $wgt(qw_i)$ is the weight of query word i, $wgt(dw_i)$

[1] excepted for definitional questions [2]

is the weight of query word i with which document word j was matched and *dist(j,questionFocus)* is the distance beetween document word j and the question focus *questionFocus*.

$$score_{passage} = score_1 + score_2 \tag{1}$$

$$score_1 = sum_{i=1}^{m} wgt(qw_i) \tag{2}$$

$$score_2 = \frac{\sum_{j=1}^{k-1} \frac{wgt(dw_j)+wgt(questionFocus)}{\alpha * dist(j,questionFocus)^2}}{k-1} * m \tag{3}$$

The main adjustments made in comparaison with [5] is the introduction of the question focus in the calculus and the consideration of the whole passage instead of only selected sentences linked by anaphora. In fact, [5] do not consider the question focus as an important element but as an ordinary element. In our application, the density is computed by mainly taking into account the distance between the question focus and the other query terms in the passage. It must be noted that by experiment the alpha value is assigned to 0.5. An other difference is that [5] compute the $score_{passage}$ coefficient for all sentences and they only gather two sentences into a same passage if for example the second sentence contains an anaphora of a noun belonging to the first sentence. The aim of their system is to provide passage which probably contains the answer. But, our application is of different nature. The density measure is used in order to rank given passage according to their ability to contain an answer to a question. So, we do not work at sentence level but at passage level.

5 Answer extraction module

After the question analysis and the passage ranking, we have to extract 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 and, on the other hand, a list of passages evaluated by our previous step according to this question [4].

This module is divided into four steps:

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).

5.1 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.

In our system, we noted that references to date and time are slightly exploited and the comparison of dates is often complicated. We thus tried to improve the recognition of dates, their standardization and their exploitation. The first stage is to locate the references to date and time. Numerical values, integers, real, and literal ones are annotated. Textual elements (days, month, etc) are also located. Then, the dates, hours and intervals of time are built.

Let's taking the following passage: "En mars 1989 , La Sept devient la Société européenne de programmes de télévision et reçoit du CSA l' autorisation d' émettre sur le satellite TDF 1 en avril 1989 ." After the labelling phase, this text becomes:

```
<duree type="date">
   <mod-pre type="eq">En</mod-pre> <mois type="car">mars</mois>
   <annee type="num">1989</annee>
</duree> , La <mois type="car">Sept</mois> devient la Société européenne de programmes de télévision et
reçoit du CSA l' autorisation d' émettre sur le satellite TDF 1
<duree type="date">
    <mod-pre type="eq">en</mod-pre> <mois type="car">avril</mois>
    <annee type="num">1989</annee>
</duree> .  Elle commence à diffuser ses programmes
<date type="lin">
    <mod-pre type="eq">le</mod-pre> <no-jour type="num">30</no-jour> <mois type="car">mai</mois>
    <annee type="num">1989</annee>
</date> ;
```

On this result, we can make a some remarks. First of all, are not regarded as "dates" only the references including a day, a month and a year. In the contrary case, this reference is labelled "duration" ("durée" in french). Moreover, we also labelled the article preceding this reference. This article gives information concerning the "direction of time" compared to the object of the sentence. Our system of temporal labelling still has some imperfections. In this example, it labels " sept " as being September whereas it corresponds rather to the integer " 7 " (the name of the television channel ; "sept" means "seven" in french). This is produced by two different processes. First of all, we have a system allowing to recognize the numerical values in form literal. Then, the months whose name is long are often shortened. Also, we parameterized the system so that it recognizes " September " but also " sept. " or " sept ". Consequently " sept " indicates an integer but also September.

To facilitate the comparison of date, we chose to calculate the elements of date (and hour) in the ISO 8601 form. For that, we have " calculated " the numerical values of the years, the months and the days. Then, we built the ISO form of the date. We made the same thing with hours. With the preceding example, we obtain then:

```
<duree type="date" iso8601="1989-03">
   <mod-pre type="eq">En</mod-pre> <mois type="car" val="03">mars</mois>
   <annee type="num" val="1989">1989</annee>
</duree> , La <mois type="car" val="07">Sept</mois> devient la Société européenne de programmes de télévision et
reçoit du CSA l' autorisation d' émettre sur le satellite TDF 1
<duree type="date" iso8601="1989-04">
   <mod-pre type="eq">en</mod-pre> <mois type="car" val="04">avril</mois>
   <annee type="num" val="1989">1989</annee>
</duree> . Elle commence à diffuser ses programmes
<date type="lin" iso8601="1989-05-30">
   <mod-pre type="eq">le</mod-pre> <no-jour type="num" val="30">30</no-jour>
   <mois type="car" val="05">mai</mois>
   <annee type="num" val="1989">1989</annee>
</date> ;
```

To go a little further, we also sought to combine date and hour. For example, the date " lundi 17 janvier 1994 13h31 " (" Monday January 17, 1994 13h31 ") will be annotated in the following way:

```
<date-time type="lin" iso8601="1994-01-17T13:31">
      <date type="lin" iso8601="1994-01-17">
            lundi <no-jour type="num" val="17">17</no-jour> <mois type="car" val="01">janvier</mois>
            <annee type="num" val="1994">1994</annee>
      </date>
      <time type="lin" iso8601="T13:31">
            <mod-pre type="eq">à</mod-pre> <heure type="num" val="13">13</heure> h
```

```
            <minute type="num" val="31">31</minute>
        </time>
</date-time>
```

This system of labelling functions but is to be improved. The objective is better to locate the temporal elements in the questions as well as in the selected passages. Unfortunately, concerning this AVE 2008 evaluation, few questions suggested were based on the temporal aspect. Moreover, the processing was undoubtedly incomplete. It thus did not produce significant results.

5.2 Named entities extraction

For locating named entities, NEMESIS tool [3] 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.

5.3 Acronym definition extraction

For acronym's definition search, we use a tool developed by E. Morin [8] based on regular expressions. It detects acronyms and links them to their definition (if it exists).

5.4 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). 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, i.e. answers extracted by a pattern with an higher priority are considered as better answers than the ones extracted by patterns with a lower priority.

As a result, for a given question, patterns associated with the question category are applied to all passages. Thus, we obtain a set of candidate answers for this question. Patterns (syntactic patterns) are based on the noun phrase that contains the focus of the question. Therefore, the first step consists in only selecting passages which could contain the answer and which contain the focus of the question.

5.5 Answer selection

When the answer type was 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 definitional questions such as the 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.

6 Validation module

The validation module is divided into two steps:

- a temporal validation
- a answer validation : comparison between AVE answer and PRODICOS answer

6.1 Temporal validation

The first step of validation module aims to compare the temporal elements of the question and the temporal elements of the question's passages. The temporal elements are calculated by the numerical entities extraction module, presented in section 5.1.

For each passage, a temporal coefficient is calculated (equal to -2,-1,0,1,2):

- If there is no temporal element in the question : score = 0 (nothing can be said)
- If there are temporal elements in the question :
  - If there is no temporal element in the passage : score = 0
  - If there are temporal elements but highly contradictory (not same year in the question and in the passage): score = -2
  - If there are temporal elements but contradictory (year not specified in the question or the passage): score = -1
  - If there are temporal elements but not completed (same year, but days or months not specified in the question or the passage): score = 1
  - If there are exactly same temporals elements in the question and the passage : score = 2

For the first time, our goal is to use the temporal coefficient to choose the best passages for the question : the passages which have a temporal coefficient equal to 1 or 2 are selected.

In the AVE 2008 task, our temporal validation module obtains the following results:

- temporal coefficient = 2 : 9 passages
- temporal coefficient = 1 : 7 passages
- temporal coefficient = -2 : 1 passages
- temporal coefficient = 0 : 182 passages

There are only 17 questions where there are temporal elements in the question and in the passages for this question. So, the temporal validation in the AVE 2008 campaign doesn't give enough informations to choose the best passages to find the answer.

Table 1: Human evaluation

Answer type	Number of answer
Validated	53
Unknown	20
Rejected	126
Total	199

6.2 Answer validation

For each question, our Prodicos system returns an answer. The answer validation aims to compare, for each question, the Prodicos answer with the answer of each passage.

The answer validation is divided in several steps:

- If the Prodicos answer is the same answer that the passage's anwswer : the passage's answer is validated
- If the Prodicos answer included in the passage's answer: the passage's answer is validated but the confidence coefficient is decreased
- The Prodicos answer is not completly included in passage's answer (depends on the number of present words of this answer) : the passage's answer is validated but the confidence coefficient is more than decreased
- Otherwise passage's answer is not validated

If there is only one passage's answer, this passage's answer is the SELECTED answer; otherwise the ProdicosAV system choose the first answer which have the best confidence coefficient as the SELECTED answer.

7 Results Analysis

French Answer Validation Exercise consists of 108 questions which can each get one or more candidate answers. There are 199 candidate answers. In this campaign, systems are evaluated in two ways. On the one hand, in order to evaluate systems, precision and recall are calculated according to all question answers. On the other hand, the second group of measures aims at comparing Question Answering systems performance with the potential gain that the participant Answer Validation systems could add to them. So, the efficiency is measured according to the answer that the system provides to an user's question.

7.1 System analysis at answers level

All answers (199) given by the system are analysed according to human judgment. It is worth noting that the "unknown" value given by a human expert to an answer is not taken into account in the evaluation. The human evaluation results are given in table 1. A first evaluation of the results obtained by ProdicosAV are given in table 2. 43 answers were validated by ProdicosAV and among them 24 are validated too by human judges. 136 answers were rejected by our system and among them 109 are rejected too by human experts. Our system obtains a precision rate equal to 0.56 and a recall rate equal to 0.46.

We made an other evaluation concerning the type of question and results obtained (table 3 and table 4).

For definitional question, the test set contains 46 answers for 29 questions. For 16 answers validated by ProdicosAV only 5 of them do not correspond to the human judgment. Therefore,

	ProdicosAV System	Human expert
Validated	43	24
Rejected	136	109

Table 2: ProdicosAV evaluation

	Validated by ProdicosAV	Validated by human	Validated by ProdicosAV and human
Definition	16	22	11
Numerical entity	7	8	3
Named entity	22	16	8
Other queries	6	7	2

Table 3: ProdicosAV evaluation

	Rejected by ProdicosAV	Rejected by human
Definition	30	15
Numerical entity	21	20
Named entity	47	47
Other queries	50	44

Table 4: ProdicosAV evaluation

the system precision is high but its recall is worse. Indeed, only 16 answers were validated by our system while 23 of them should have been. The problems come mainly from question analysis problem (for example question 188: "Vasa" tagged as verb), from pattern extraction problems (for example question 159: "Jane Austen (16 décembre 1775, Steventon, Hampshire - 18 juillet 1817, Winchester) est une écrivain") or from acronym extraction problems (the meaning of the acronym is translated in the passage what implies that acronym's letters are completely independant of it).

For numerical entity question, the test set contains 28 answers for 16 questions and for named entity question, the test set contains 69 answers for 36 questions. Only 11 answers of 29 answers validated by ProdicosAV correspond to the human judgment. And the system recall is also worse, indeed only 11 answers were validated by our system while 24 of them should have been. For other questions, the test set contains 56 answers for 27 questions. The system precision and rappel are equivalent : only 2 answers of 6 answers validated by ProdicosAV and of 7 answers validated by the human judgment. The problems come mainly from the selection of the candidate passages : the question focus doesn't detect in a lot of passages or from question analysis problem or from reference's absence.

7.2 System analysis at questions level

The second group of measures aims at comparing QA systems performance with the potential gain that the participant Answer Validation systems could add to them. The test set includes 108 questions. Human experts find a response for 52 of them. Among them, our system gives 23 responses which are well validated. Human experts gives a negative response for 56 questions, among them, our system gives 38 negative responses. The *qa-accuracy* rate obtained is 22%. The maximum that the system should have obtained is 48% (according to the number of validated response the humans found). The *qa-rej-accuracy* rate obtained is 35%. The maximum that the system should have obtained is 52% (according to the number of rejected response the humans found). Our system obtains a not satisfactory *estimated-qa-performance* rate equal to 29%. The maximum that it should have obtained is 73%. This shows that ProdicosAV System has not a

good ability to acknowledge the identification of questions with a set of answers in which no correct one has been found.

8 Conclusion and Prospects

The results are not satisfactory, because we only recover 24 correct answers on 53 correct answers. The problems come mainly from question analysis problem (specialy the word's labeling), from pattern extraction problem (specialy the absence of semantics and coreference). We can also improve our validation module by taking into account all the answers proposed by the various strategies of extraction module and not only the best. We can also take into account the another external informations as the passage's date, etc.

References

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

[2] Monceaux L. : "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)

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

[4] Desmontils E., Jacquin C., Monceaux L. : "Question Types Specification for the Use of Specialized Patterns in Prodicos System", in C Peters, F.C Gey, J Gonzalo, G.J Jones, M Kluck, B Magnini, H Müllern et M De Ruke eds. Proceedings of Accessing Multilingual Repositories, 7th workshop of the Cross Language Evaluation Forum, CLEF 2006, Revised selected papers, volume 4730 de Lectures Notes of Computer Sciences (LNCS), Springer-Verlag Berlin Heidelberg , pp. 280-289, 2007.

[5] Lee G. G., Seo J., Lee S., Jung H., Cho B., Lee C., Kwak B., Cha J., Kim D., An J., Kim H., Kim K.: "SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP". Proceedings of tenth Text REtrieval Conference (TREC 2001), 2001.

[6] S. Tellex S., Katz B., Lin J., Fernandes A., Marton G.: "Quantitative evaluation of passage retrieval algorithms for question answering", In SIGIR conference on Research and development in information retrieval, pages 41–47. ACM Press, 2003.

[7] L. Monceaux, C. Jacquin, E. Desmontils : The query answering system Prodicos, in C Peters, F.C Gey, J Gonzalo, G.J Jones, M Kluck, B Magnini, H Müllern et M De Ruke eds. Proceedings of Accessing Multilingual Repositories, 6th workshop of the Cross Language Evaluation Forum, CLEF 2005, Revised selected papers, Vienna, Austria, september 2005, volume 4022 de Lectures Notes of Computer Sciences (LNCS), Springer-Verlag Berlin Heidelberg , pp. 527?534, 2006

[8] E. Morin, "Extraction de liens sémantiques entre termes á partir de corpus de textes techniques", PhD Thesis, Université de Nantes, LINA, De'ecembre 1999. http://www.sciences.univ-nantes.fr/info/perso/permanents/morin/article/morin-these99.pdf