MIRACLE at the Spanish CLEF@QA 2006 track

César de Pablo-Sánchez, Ana González-Ledesma,
Antonio Moreno, José Luis Martínez-Fernández, Paloma Martínez
Universidad Carlos III de Madrid
{cesar.pablo,paloma.martinez}@uc3m.es
Universidad Autonoma de Madrid
{ana,sandoval}@maria.lllf.uam.es
DAEDALUS S.A. - Data, Decisions and Language, S.A.
jmartinez@daedalus.es

Abstract

We describe the prototype QA system built by MIRACLE group, a group composed by three Madrid universities and the spin-off, DAEDALUS. The system is an elaboration of our last year system with several improvements in question analysis and NERC components. We submitted two runs for the Spanish runs with different strategies to use NE in passage selection and answer ranking. Results show that a recall oriented approach obtain more accurate results. A detailed analysis of errors and a preliminary comparison with our last year system are also discussed.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids

1 Introduction

In our third participation to the CLEF@QA task, the MIRACLE group submitted two runs for the Spanish monolingual subtask. Runs differ in the way they consider Named Entities and other complex expressions in passage selection and ranking. The core of the system is taken from our last year's effort[1] although it features improvements in most of their modules.

The system has also been adapted at an extent to new requirements of 2006 guidelines for the QA task. Questions no longer are provided with a question type and the system should infer this from the question text. New questions types are also introduced, in particular list of factoids. Although our system has been prepared to categorize questions according to these question types, not all of the types have received the same attention due to time constraints. In particular, lists are treated as simpler factual questions. Other novelties in the task include the normalization of timexes which we have also accomplished partially. These improvements also allow our system to permform a more elaborate treatment of temporally restricted questions. Finally, the guidelines also required short passages supporting the answer and several answers(up to 10), although, in the end, only one pair *(answer,snippet)* has been oficially evaluated.

2 System description

The system is organized in a pipelined architecture where we have the three classic main modules, Question Analysis, Passage Retrieval and Answer Extraction. These modules use common language analysis services for Spanish. We have also developed a tool that logs final and intermediate results and allow further analysis of the system using a web interface.

2.1 Language analysis

Our language analysis module is composed of DAEDALUS STILUS[6] analyzer service and some other tools we have developed specially for these task.

STILUS provides tokenization, sentence detection and token analysis for Spanish. The analysis of tokens include information about POS tags, lemmas and other morphological features such as number and tense. The analysis is also enriched with semantic information which is stored in a dictionary of common Named Entities (NE) extracted from several resources. Those NE are organized following Sekine's typology[5]. Themathic relations and geographical relations for some of those NE are also given although we have not used them in the QA system yet. These tool has been developed for spell, grammar and style checking in mind and therefore is very exhaustive in language coverage and their analysis, even if some are infrequent uses. As this tool is reused for this task it does not include a proper POS tagger or chunking like those found in other text analysis toolkits more focused in similar tasks. In contrast it has excellent performance for large quantities of text and is fast, stable and robust.

To adapt the output of STILUS to the processing requirements of the QA task we include some more language analysis modules. A very basic set of rules remove infrequent uses of some common words based on contextual information. To improve recognition of NE, in particular recall, another set of rules are used. This step recognizes and groups tokens based on the previous information and other ortographic, contextual and structural features. Most semantic information is retained and even some rules could add new entity class information to the analysis. This module has been revised and improved since last year submission.

Finally, for Temporal Expressions (TE) an additional normalization step is also performed. This feature is in current development but our intention id that allow our system to develop a shallow temporal reasoning method.

2.2 Question analysis

This module has been extended to deal with new kind of questions and the fact that the question type would not be given as an explicit input to the systems.

The question analysis module is responsible for transforming the question string into a common representation that could be used throughout the system. Our question model, the basis for question representation, is very simple as we lack complex analysis resources and we also believe that a simpler question model is more adequate for cross-lingual applications. We characterize each question with the following features as it has been proposed in earlier work [2]: question type (QT), expected answer type (EAT), question focus (QF), answer type term, query terms and relevant terms. Query terms are those considered to retrieve candidate documents and are used to build queries. On the other hand, relevant terms are a broader set of terms that could help to locate an answer but would retrieve noisy documents as it often happens with the answer type term.

Before classification, the question text is analyzed using the language analysis services described above. Question classification is carried in two steps. Firstly, question is classified regarding their question type and later on the expected answer type is assigned. We have used a set of handwritten rules to perform question type classification. Four question types are detected, Factoids (F), Temporally Restricted Factoids (T), Lists (L) and Definitions (D). Besides common patterns, the classification use specific information for any of the types. For instance. the detection of TE usually signals for temporally restricted questions and the use of plurals in answer type terms or questions words allow us to detect list questions.

The expected answer type is also assigned using rules and lists of related words generated by a linguist. This year we improved coverage for this classification by adding more patterns and extending the lists. F,L and T questions shares the rules and the same hierarchy of answer types used last year[1]. For D questions the type reflects the object to describe or action to take (Person, Organization, Other or Acronym expansion) and uses mainly the information of our NERC module.

2.3 Document retrieval

As last year we used Xapian[7] for the retrieval stage. Xapian is an open source engine based on the probabilistic retrieval model that includes the Okapi BM25 scoring funtion [4]. The collection has been indexed using simple terms stemmed with Snowball stemmer[3] for Spanish and query terms are processed in a similar fashion.

Some improvements have been done in the query generation step. Simple terms are used in the query and composed with OR operator to improve recall. Complex terms like NE are decomposed and joined using a relaxed AND operator (AND_MAYBE) implemented by Xapian. With this operator, documents that contained all the component terms are ranked higher, but if some of this terms does not appear documents are not removed from the result set.

Documents selected by the retrieval step are analyzed and sentence boundaries are idenfied. Snippets that contain a number of relevant terms from the questions are considered for further processing. The threshold to select a snippet is proportional to the number of relevant terms that represent the questions and in particular if it contains the entity in focus. So far, we have only considered snippets that spans a single sentence.

2.4 Answer extraction

To perform answer extraction a filter for every expected answer type in our system has been developed. Answer extraction used this information to select candidate answers from the set of relevant sentences previously identified. For every pair of candidate answer and snippet a score is given, considering factors as the document score and the frequency and ratio of relevant terms contained. These scoring step that we call local score consider only information present in one snippet.

In a second step, similar candidate answers are conflated and a new global step is generated. Rules for temporal expression normalization has been developed this year and used to conflate dates. Other simple techniques that filter stopwords and compare the rest of the tokens are used for other NE types. Those techniques are rather heuristic and definitely does not conflate different variations of a name but have been proven useful. When grouping several candidate answers, the one with the higher score is selected as representantive for that group, together with its snippet. A redundancy score is calculated by counting the number of snippets in a group up to a maximum of *N*. A global score is assigned to the group that combines the local score (*wls*) and the normalized redundancy score (*wrs*) to produce a confidence between 0 and 1. The values used in our experiment has been assigned manually and are *N=10*,*wls=0.9*,*wrs=0.1* .

3 Description of the runs

Two runs were submitted to the Spanish monolingual task. they differ in the way that use multiwords or complex terms. Complex terms includes multiword units such as most NE, TE and

Table 1: Evaluation results for MIRACLE's submitted runs

Run	*R*	*X*	*U*	*Acc*(%)	*Acc*(*F*)(%)	*Acc*(*D*)(%)	*Acc*(*T*)(%)	*CWS*	*K*1
mira061eses	36	3	6	18.95	19.86	16.67	0.00	0.05708	-0.3405
mira062eses	40	4	7	21.05	20.55	23.81	0.00	0.07896	-0.3977

numerical expressions. As we index the EFE collection using simple terms it is not clear which strategy works best for selecting and scoring relevant documents and sentences. Before thinking of more complex alternatives, we have tried different simple alternatives in each of the runs. Both runs decompose complex proper names in a query into simple terms joined with the operator AND_MAYBE to retrieve documents.

Run mira01eses selects and scores sentences considering the whole multiword. This strategy select those sentences in which the expression is the same that appear in the question and is clearly oriented to favour high precision. In contrast, run mira02eses uses terms that compose the term. This run would favour recall because it could match diffrent common forms of referring the same entity for example, but it is also problematic as it could select noisy sentences that refer to other entities.

4 Results

Table 1 presents the official results for factual (F), definition (D) and temporally restricted questions (T) for the two monolingual spanish runs submitted. Despite there are several questions that has temporal restrictions (33 in our intrepretation), only two of them have been officially assigned this question type. Neither of them have been answered correctly.

General performance for factual and definition questions is quite similar for both runs while it seems that mira062eses obtain better accuracy score and CWS. The interpretation of the results need further investigation but it seems that the precision oriented run (mira061eses) filter some useful sentences. This is also supported by the higher number of NIL responses in this run.

In contrast, K1 measure, that takes into account the self score of the system, it is better for run mira061eses than mira062eses. The difference in the scoring for the two runs is motivated by considering complex terms as a unit or separated. Although, both methods perform rather bad with regard to this metric, the second seems to be worse as it usually produces higuer scores for the same sentence and threfore higher confidence. In any case, the inspection of the results suggest that correct and incorrect candidate answers are scored too close. This is also, in our opinion, the source of errors for unsupported answers as it is probably produced for similar scores before or after answer normalization.

For questions that we have believed that have a temporal restriction, we have performed an unofficial evaluation. Both runs have similar overall results with accuracy figures comparable to usual factoid questions. If we consider some doubtful cases (unsupported answers and some cases that obtain all the same score) this percentage is even higher. It seems that our effort to process temporal information are in a good direction even if there is in an initial stage. This is also supported from the fact that the system obtain good scores for F-TIME questions.

In addition, the 2006 exercise have run a test on list questions. The results for our system have been very low. Run mira061eses answered only 2 questions right, with a P@10 score of 0.03 (three correct answers). The second run (mira062eses) was even worse with only one right answer. Although list questions have been identified in the question analysis phase, no special processing has been implemented for them due to time constraints and they are treated as simple factoids.

Table 2: Error analysis for run mira061

Error type	%
Question classification	7,79
Question analysis	18,18
Document retrieval	38,96
Sentence selection	3,89
Answer extraction	11,03
Answer ranking	20,12

Table 3: Error comparison between 2005 and 2006 runs

Error type	2005	2006
Questions classification & analysis	25.98	25.97
Document retrieval	20.81	38.96
Sentence & answer extraction	11.83	14.92
Answer ranking	40.84	20.12

4.1 Error Analysis

It is important to estimate the causes for low scores got this year, analysing which module in the general system is the main responsible for the errors. Of course errors in a QA system cannot be assigned to only one system, and this should be beared in mind when judging the following error analysis.

With respect to our 2005 error classification, we added two new types this year, to better understand the treatment of the questions and answers in our system. The estimation of errors for run mira061eses are shown in the Table 2. In the first column, we show the percentage of error types with respect to the total of errors considering Wrong, ineXact and Unsupported (in the case of mira061eses, 154 questions).

Comparing the error classification in both years (see Table 3), we can observe significant differences in the approaches. While question analysis and answer extraction (although with addings and modifications) have obtained basically the same results in both years; in document retrieval and answer ranking the results have been reversed. A possible explanation of this year bad results in document retrieval is the increasing difficulty of the questions. We should advise that this is a preliminary study, as last year we had correct assesments for all the questions before the workshop, while these year we are considering only our judgments. Definitive results would be presented in a final version of this report.

Besides, careful inspection of the results have allowed us to detect at least two significant bugs. The first of them affect the classification of some definition questions as *¿Qué es el CERN?* that should expand and acronym. The other bug appears in the extraction of candidates for some NE types.

5 Conclusion and Future Work

In these work we have studied how to consider complex terms or multiwords in the selection and ranking of answers in our Question Answering system for Spanish. Results show that favouring recall by dividing multiwords and considering single terms helps to locate more candidate sentences and that improves accuracy. The two runs submitted explore different ways to consider multiwords that lie at different extremes. Using multiwords as a single unit seems to help to identify NIL answers which could be interesting in some practical applications. A technique that combine evidence from both approaches could obtain much better results. Different alternatives consist on mixing the two ranking list or extending analysis to consider simple coreference at the document level. Further investigation on the correct way to weight the evidence of the terms for ranking and self-scoring is still needed as the result of different measures are contradictory. We probably need to consider again factors as distance and weight that we already used last year and that this year were disregarded. Another issue that need to be solved it is the proper scaling of the confidence score to be more informative.

In contrast regarding the recognition and normalization of temporal expression it seems that we are in the good track. We plan to continue this work that we espect that would allow to do some shallow reasoning for TIME questions and temporally restricted questions.

Finally, it seems that more effort should be put to other aspects of the system to reach the desirable performance. We have other open lines of work regarding the use of relation extraction patterns for some question types and improving retrieval with the experience acquired in other evaluation tasks.

References

[1] de Pablo-Sanchez C. et al. Miracle's 2005 approach to cross-lingual question answering. In *Working Notes for the CLEF 2005 Workshop. Vienna,Austria*, 2005.

[2] Marius Pasca. *Open Domain Question Answering from Large Text Collections*. CSLI Publications, 2003.

[3] Martin Porter. Snowball stemmers and resources website. On line http://www.snowball.tartarus.org, July 2006. last visited.

[4] S.E. et al. Robertson. Okapi at trec-3. In D.K. Harman, editor, *In Overview of the Third Text REtrieval Conference (TREC-3)*, 1995.

[5] Satoshi Sekine. Sekine's extended named entity hierarchy. On line http://nlp.cs.nyu.edu/ene/, August 2006. last visited.

[6] Stilus website. On line http://www.daedalus.es, July 2006.

[7] Xapian: an open source probabilistic information retrieval library. On line http://www.xapian.org, July 2006. last visited.