MIRACLE's 2005 Approach to Cross-Lingual Question Answering

César de Pablo-Sánchez[1], Ana González-Ledesma[2], José Luis Martínez-Fernández[1,4], José Maria Guirao[3], Paloma Martinez[1],Antonio Moreno[2]

[1] Universidad Carlos III de Madrid
[2] Universidad Autónoma de Madrid
[3] Universidad de Granada
[4] DAEDALUS - Data, Decisions and Language, S.A.

{cesar.pablo,paloma.martinez}@uc3m.es, {ana,sandoval}@maria.lllf.uam.es, jmguirao@ugr.es, jmartinez@daedalus.es

Abstract

This paper presents the 2005 MIRACLE's team approach to CLEF QA with Spanish as a target task using miraQA system. The system is based on answer extraction and uses mainly syntactic patterns and semantic information. Six runs were submitted for Spanish, English and Italian as source languages using commercial translation software. The system performs reasonably well for Spanish factual questions if compared with other participants but its performance is lower with definition and temporally restricted questions. A thorough error analysis has been carried out to spot critical points for improvement. Comparison of cross-lingual runs shows that sometimes, for the cross-lingual task, answers are found that, for the monolingual tasks, cannot be located or do not appear as the first option.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval ; H.3.4 Systems and Software. E.1 [Data Structures]; E.2 [Data Storage Representations]. H.2 [Database Management]

Keywords

Linguistic Engineering, Information Retrieval, Question Answering, Cross-Lingual Information Retrieval.

1 Introduction

Question answering systems localize and extract concrete answers to information needs expressed in natural language, usually in the form of questions. Information can be stored in different ways, from structured databases to unstructured document collections but still natural language is a convenient or preferred code for lots of users to access it. Besides, we do not need to assume that the information demanded by the user is in his or her own language, and therefore the issue of breaking the language barrier also arises. This seems an important issue in some applications of QA systems in domains like turism but also for accesing information in the web.

This paper presents and analyzes the results of our second participation in the CLEF-QA task. We have submitted six runs with Spanish as a target language, but with different source languages, Spanish, English and Italian. Our system, miraQA, is mainly based on answer extraction and uses low level linguistic analysis. In contrast, we have incorporated some semantic resources for NE recognition. The approach and tools are different from last year [9] but we believe that both could be combined in a near future. Runs use different strategies for answer extraction and selection. Cross-lingual runs use direcct translation of the questions from source to target language. A further inspection of the errors made by the system in mono and cross-lingual runs has been carried out and, as a result, some ideas for further improvement and their priority are also presented.

2 Description of the MIRACLE Toolbox

MIRACLE's contribution to CLEF QA 2005 is an almost new development based on the experience acquired after last year initial contribution. The system, miraQA, uses different individual resources from MIRACLE's toolbox as well as open source components. Our aim is to achieve an architecture where we could easily plug in different resources for comparison and (semi) automatic evaluation of the system and their different parts. The

number of resources produced by CLEF, especially MultiSix [6] and MultiEight corpora, allows us to put in practice an “agile” development methodology that help us to evaluate frequently.

The system is based on a classical pipelined architecture, what we call a “U” architecture as presented in Figure 1.

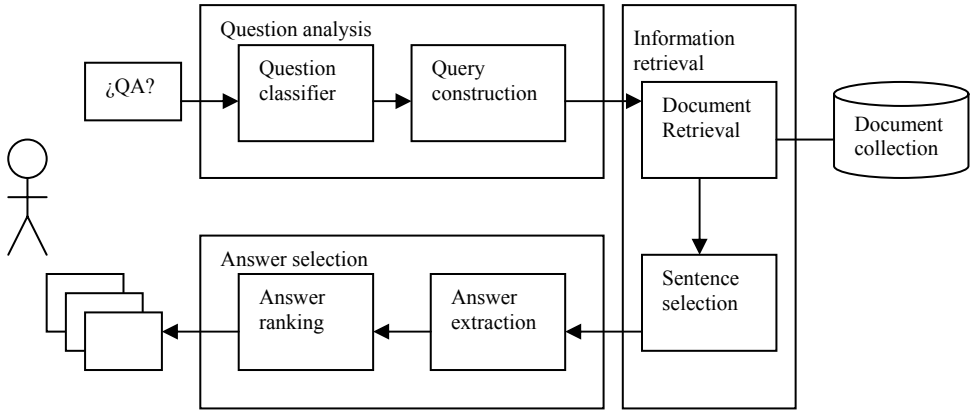

Figure 1: miraQA "U" logical architecture

The resources that we have used in the system are:

- STILUS©[1] linguistic processors, in particular a morphosyntactic processor. This tool was initially developed for spell and grammar checking. It produces all possible morhopological analysis for a word and assigns all possible tags using a large dictionary of Spanish. The tool also contains a large dictionary of common collocations and recognizes and normalizes some usual complex tokens for commercial applications as dates, money and other numerical expressions and web addresses. Besides, it has been extended to recognize Named Entities of different kinds like person names (first and last), countries, cities and other geo-political entities, nationalities, organizations, etc. Recognized entities are tagged following Sekine´s taxonomy [12].
- Xapian [15]. As last year we are using this probabilistic information retrieval engine to index and query the collection of EFE documents. Xapian provides ranks based on the Okapi BM25 model.
- Machine translation. For the cross-lingual experiments we used Systran[11] to translate questions in Italian and English. Questions were passed to the Spanish miraQA system without any further processing.

Next sections describe the details of the logical components of the architecture. The same basic process is applied to the different types of questions and, in particular, temporal questions are considered as factual ones.

2.1 Question taxonomy

The Questions taxonomy in miraQA is determined by answer types. As most of the factoid questions are expecting some kind of named entity we initially consider Sekine’s NE taxonomy, which contains more than a hundred NE types hierarchically linked. Most of the types are too specific for the tools and resources that we have available. So we actually prune the taxonomy to consider only NE types that could be extracted with some confidence. In contrast, we added some answer types like acronyms or a particular kind of definition (short descriptions that are often realized as appositives) that we called properties. The taxonomy preserves the hierarchy as we think that this could provide a way to back off to larger semantic classes of NE types if more specific answers are not found.

2.2 Question classification and query generation

Question classification is achieved using a set of linguistic rules produced after some deep study and generalization of CLEF 2004 Spanish data. Question classification aims to assign the correct answer type depending on the category of the question. Definition questions mainly ask for organization names or for short descriptions of persons or organization so this are the two types considered. For factual and temporal questions the taxonomy presented above is used. The classification proceeds in three steps, (1) question is analyzed using

[1] More information about STILUS linguistic tools for Spanish can be found at www.daedalus.es

STILUS©, (2) feature extraction based on some simple heuristics and (3) proper classification. The features used to classify the questions are the following:

- IsInterrogative
- InterrogativeLemma
- IsNounBeforeVerb
- FirstNounLemma (Question focus)
- MainContentVerbLemma
- FirstEntityType

Most of the interrogatives very often determine the type of the questions, while for Spanish, at least *Qué* y *Cuál* pronouns are used for almost every question type. For questions with this interrogatives the Question focus is the most determinant word and for that reason we have compiled lists of common question focuses for each type. The output from the processor is ambiguous and for that reason some simple rules to disambiguate verbs and nouns for feature extraction are used.

Rules that classify question have been manually generated from inspection of past CLEF campaign. So far, we have manually tagged the 200 hundred ES-ES questions with Sekine answer types and question form tags(affirmative/interrogative/relative) at the question level. Inside sentences we have tagged POS tags, Named Entitities, shallow syntactic structure and semantic arguments. Our objective is to achieve a multilevel tagged corpus of questions where deeper question structure could be inferred or tested. Our experience shows that available resources for Spanish are more inaccurate analyzing questions than typical sentences from documents.

After that step, queries to be passed to the search engine are directly formed with relevant terms extracted from the questions. Some terms are believed to harm retrieval effectivenes as they are too common in text and usually produce many noisy documents. A second specific stopword list, mainly composed of question focuses for the specific answer type, is applied to filter terms used in the query. In contrast, these terms are used later on for sentence filtering.

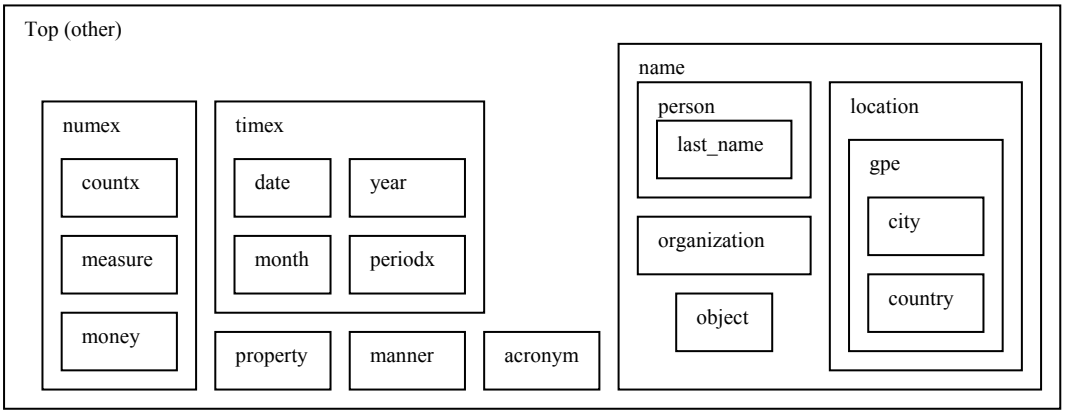

Figure 2: Question taxonomy

2.3 Document and Sentence Retrieval

Documents are indexed off-line using Xapian to create indexes. The submitted runs used the typical text operations of stopword removal and stemming as provided by Xapian engine, so the Snowball [10] stemmer for Spanish was used.

At retrieval time, the first N results returned by the engine are analyzed using STILUS© tools and sentences are filtered and scored according to the number of content terms in common with the query. Sentences with less different content terms than a threshold are discarded although this threshold is 1 for queries that have fewer content terms. The results from our runs were produced from the first 100 retrieved documents, as we have experimentally tested that few documents contain candidate answers after this limit.

2.4 Answer extraction and selection

Answers are extracted using rules depending on the answer type identified by the question classifier. An specialized answer recognizer is written for every category using a kind of automata that evaluates boolean predicates over annotated tokens. As the output from the STILUS© processor is not disambiguated, rules are robust enough to deal with some common problems. For most entity names these rules try to group recognized subunits in order to improve recall. Predicates used in the automata check for orthographic, morphological and syntactic and semantic features.

For some answer types like MANNER and OTHER, it is difficult to establish a model of the answer and for that reason in our second run we tried an even simpler extractor. Candidates in this extractor are selected as chunks between content words.

After extraction, similar extracted answers are conflated using some simple rules that remove stopwords and some spurious content words from the query. For every group of answer the system picks as representative the answer with higher score, which will provide the source document.

MiraQA scores every answer in two steps. Runs 051 scored sentences according to the inverse frequency of relevant terms appearing in the query. Answer instances were given the same score than the sentence. Runs 052 used a weighted combination of tf*issf (inverted selected sentence frequency) terms and median distance from keywords to answers.

In a second step instances are conflated into groups and the answer with the best score is elected as representative. Redundancy is considered by computing the linear combination of the score and the ratio of documents that support an answer group.

3 Basic experiments

We have submitted two runs for three language pairs. The target language for all of them is Spanish and the source languages are Spanish, English and Italian. The evaluation measures for the runs are presented in Table 1. Runs differ in the ranking function used to order answers and in the strategy used to answer OTHER questions. For the runs numbered 051 OTHER questions extract mainly noun phrases while in 052 runs chunks of words between keywords are extracted.

Best results were achieved in mira051eses run but differences do not seem significant when only the first ranked answer is considered, while the runs contained a moderate number of different answers. In fact for English and Italian runs, the number of correct results are almost the same, being the figures of the weighted evaluation measures different. As could be expected, accuracy is lower for cross-lingual runs with a loss between 5% and 7%.

The system processes temporal questions in a similar way to factual questions and the accuracy obtained for the former ones is much lower than for the latter ones. The increasing effect in accuracy for English temporal questions is in fact due to answer with correct NILs. The system performs better for definition questions than for the rest of types in absolute numbers. In contrast, compared to other systems with Spanish as a source language, miraQA is answering better factual questions, in particular questions of the PERSON class.

Run	R	X	U	Acc.	Acc(F)	Acc(D)	Acc(T)	CWS	K1	Corr
Mira051eses	**51**	**11**	**0**	**25,5**	**26,27**	**34**	**9,38**	**0,12372**	**-0,3021**	**0,3157**
Mira052eses	46	14	0	23	22,03	34	9,38	0,10382	-0,3432	0,316
Mira051enes	39	7	1	19,5	16,95	28	15,62	0,09376	-0,3922	0,23
Mira052enes	39	8	2	19,5	16,95	28	15,62	0,08809	-0,3943	0,2278
Mira051ites	36	10	0	18	16,95	26	9,38	0,06829	-0,4379	0,2244
Mira052ites	35	11	0	17,5	16,95	24	9,38	0,07186	-0,4471	0,2192

Table 1: Statistics for assesed results

A deeper error analysis have been carried out in order to characterize the performance of the modules, detect problems and assign error rates. We have computed the classification accuracy according to our own question taxonomy. Results show that the accuracy for Spanish questions in CLEF 2005 test set is 80,50% being the main source of errors due to the lack of coverage of the words lists used to compare with question focus.

For English questions the accuracy is 77% while for Italian the accuracy is much lower (63,50%). While the source of problems for the English test set is mainly the same, some new errors are introduced by the translation

engine changing the usual order of the question. The much lower performance in Italian is due to the incorrect translation of question word “Qué?” por “Cuál?” and the lack of appropiate rules.

The second point where we have measured error is after document retrieval and results are presented in Table 2. We have used the judgments for all of the systems to compute the number of questions for which a document with and answer is retrieved. We have used documents whose answers are assessed as correct (R) or inexact (X) to compute the number of questions that have a potential answer at a certain document cut. The measure should be taken as a lower bound as for the questions that none of the systems answers correctly, we do not have an associated document. Besides, some more documents could contain correct answers that are not identified. The loss in performance for cross-lingual experiments is mainly due to errors in the lexical construction selected by the automatic translator. Documents are ranked in the order provided by Xapian which reflects that is a reasonably good feature for answer ranking purposes.

	ES	EN	IT
A@20	94	80	79
A@40	115	100	101

Table 2: Analysis of retrieved documents

	ES
A@1	44
A@2	62
A@5	81
A@10	89
A@20	94
A@40	99

Table 3 : Analysis of correct answers

Table 3 shows the manual judgement of correct answers for run mira052eses at different numbers of possible answers. The conditions are the same than before so they should be taken as a lower bound. The results indicate that the maximum performance for questions with at least an answer, even with perfect ranking, would be of 55%. The ranking function works reasonably well but there is room for improvement.

Errors in a QA system cannot be assigned to only one subsystem, as there are usually complex interplays between the different parts. For example, a low precision in the answer extraction module will make the task of the answer selection module most difficult and therefore more prone to errors. Table 4 shows an estimation of the errors of the different modules based on the measures above.

Module	Error resp.
Question analysis errors	25,98%
Document retrieval recall errors	20,81%
Answer extraction recall errors	11,83%
Answer selection errors	40,84%

Table 4: Estimation of error responsibility

Finally, we have detected that for some questions, the answer is found in cross-lingual runs while monolingual runs fail to provide a correct answer at least as a first choice. We have analyze judgements between mono and cross-lingual runs in order to quantify this performance. Changes from right to wrong in cross-lingual runs are mainly due to the incorrect translation of Named Entities, especially acronyms in definition questions. Another source of errors is the incorrect classification of questions and the incorrect translation of terms. In contrast, we believe the change from wrong to right is due to the use of different lexical alternatives at translation and their interplays with the retrieval and ranking systems.

mono	cross	ES -> IT	ES -> EN
R	U	0	1
R	W	22	17
X	W	6	5
X	R	0	1
W	X	5	2

W R 7 5

Table 5 : Answer change between mono and cross lingual runs

4 Future work

Results from the previous sections suggest that performance could be easily improved by means of using better answer selection techniques. Answers are found in 55% of the questions but only ranked in first position in half of them. We believe that better ranking functions and candidate answer filters in the style of our CLEF 2004 submission would help us in a future. Ranking functions should be suitable not only for ordering different candidate answers but also be informative enough for the user, reflecting the confidence that the system can assign to an answer.

Further work need also to be done in question classification in order to reduce their influence in later stages. We plan to include Wordnet and related Euro-Wordnet as well as the use of robust machine learning techniques in order to reduce the influence of incorrect translations.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, by means of the project RIMMEL Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01.

Special mention to our colleagues of the MIRACLE team should be done (in alphabetical order): Ana María García-Serrano, José Carlos González-Cristóbal, José Miguel Goñi-Menoyo, Sara Lana-Serrano, Ángel Martínez-González and Julio Villena.

References

[1] University of Neuchatel. page of resources for CLEF (Stopwords, transliteration, stemmers, …). On line http://www.unine.ch/info/clef/. [Visited 13/07/2005]

[2] S. Abney, M. Collins, and A. Singhal. Answer extraction. In ANLP-2000, 2000.

[3] Goñi-Menoyo, José M; González, José C.; Martínez-Fernández, José L.; and Villena, J. MIRACLE's Hybrid Approach to Bilingual and Monolingual Information Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491, pp. 188-199. Springer, 2005 (to appear).

[4] Goñi-Menoyo, José M.; González, José C.; Martínez-Fernández, José L.; Villena-Román, Julio; García-Serrano, Ana; Martínez-Fernández, Paloma; de Pablo-Sánchez, César; and Alonso-Sánchez, Javier. MIRACLE's hybrid approach to bilingual and monolingual Information Retrieval. Working Notes for the CLEF 2004 Workshop (Carol Peters and Francesca Borri, Eds.), pp. 141-150. Bath, United Kingdom, 2004.

[5] Goñi-Menoyo, José Miguel; González-Cristóbal, José Carlos and Fombella-Mourelle, Jorge. An optimised trie index for natural language processing lexicons. MIRACLE Technical Report. Universidad Politécnica de Madrid, 2004.

[6] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, M. de Rijke, The Multiple Language Question Answering Track at CLEF 2003. (see chapter "Gold Standard for the Cross-Language Tasks"), in Carol Peters, editor, Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway, 2003.

[7] Martínez, José L.; Villena, Julio; Fombella, Jorge; G. Serrano, Ana; Martínez, Paloma; Goñi, José M.; and González, José C. MIRACLE Approaches to Multilingual Information Retrieval: A Baseline for Future Research. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 210-219. Springer, 2004.

[8] Martínez, J.L.; Villena-Román, J.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.): Evaluation of MIRACLE approach results for CLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

[9] de Pablo, C.; Martínez-Fernández, J. L.; Martínez, P.; Villena, J.; García-Serrano, A. M.; Goñi, J. M.; and González, J. C. *miraQA*: Initial experiments in Question Answering. Working Notes for the CLEF 2004 Workshop, pp. 405-411 (Carol Peters and Francesca Borri, Eds.), pgs. 371-376. Bath, United Kingdom, 2004.

[10] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org. [Visited 13/07/2005]

[11] SYSTRAN 5.0 translation resources. On line http://www.systransoft.com. [Visited 13/07/2005]

[12] Sekine, Satoshi. Sekine's Extended Name Entity Hierarchy. On line http://nlp.cs.nyu.edu/ene/ . [Visited 18/08/2005]

[13] Villena, Julio; Martínez, José L.; Fombella, Jorge; G. Serrano, Ana; Ruiz, Alberto; Martínez, Paloma; Goñi, José M.; and González, José C. Image Retrieval: The MIRACLE Approach. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 621-630. Springer, 2004.

[14] Villena-Román, J.; Martínez, J.L.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.); MIRACLE results for ImageCLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

[15] Xapian: an Open Source Probabilistic Information Retrieval library. On line http://www.xapian.org. [Visited 13/07/2005]