AliQAn and BRILI QA systems at CLEF 2006[*]

S.Ferrández[1], P.López-Moreno[1], S.Roger[1,2], A.Ferrández[1], J.Peral[1], X.Alvarado[1], E.Noguera[1] and F.Llopis[1]

[1] Natural Language Processing and Information Systems Group

Department of Software and Computing Systems

University of Alicante, Spain

[2] Natural Language Processing Group

Department of Computing Sciences

University of Comahue, Argentina

{sferrandez,sroger,antonio,jperal,elisa,llopis}@dlsi.ua.es

P.Lopez@ua.es

Abstract

An initial version of AliQAn participated in the CLEF-2005 Spanish Monolingual Question Answering task. AliQAn has been improved and it has better results than last year. The system offers new and representative patterns in question analysis and extraction of the answer. A new ontology of questions types has been included. The inexact questions have been improved. The information retrieval engine has been modified considering only the words for the question analysis module as well as we do not use the question expansion of IR-n. Besides, many MACO and SUPAR errors have been solved and finally, dictionary about cities and countries have been incorporated. Finally, the BRILI system is presented as an extension of the AliQAn system, in order to deal with bilingual tasks. The final results achieved after that we improve our approach (overall accuracy of 37.89% for monolingual and 21.58% for bilingual) are shown in this paper.

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 Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids

1   Introduction

AliQAn is a monolingual open-domain Question Answering (QA) system has been developed at the University of Alicante for three years ago. Last year, AliQAn participated in the CLEF-2005[1] competition and this year, it has participated in the CLEF-2006[2] with a new version of our system. On the other hand, an extension of AliQAn for bilingual task is presented: BRILI system. This approach is able to answer English questions from Spanish documents.

Both systems are fundamentally based on the use of syntactic patterns for the identification of candidate answers, but in the case of AliQAn the set of patterns has been extended. Word Sense Disambiguation (WSD) is applied to improve the system. WSD has several critical problems like the running time of WSD algorithms with huge corpora required to QA systems and the low precision of WSD algorithms. In order to solve these problems, we applied a concrete WSD algorithm [4] that reduces its running time in 98.9%. Besides, there are many modifications that have allowed the construction of a system more robust. The patterns design has been improved, dictionaries have been incorporated in the extraction of the answers, a treatment to solve the inexact answers and errors of MACO [1] and SUPAR [2] have been resolved. This article focuses on the improves have been applied on our system that we are going to explain in the content of this paper.

The rest of this paper is organized as follows: section two the structure and the functionality of AliQAn and its main components are detailed. Section three describes our participation in a Cross-lingual task with the system BRILI. Afterwards, the achieved final results discussed in section four and finally, section five explains our conclusion and future work.

2   System Description

2.1   Overview

In this section, the structure and functionality of AliQAn are detailed. Some new characteristics and improvements in our system are described with respect to the architecture used in CLEF-2005 [6].

AliQAn is based fundamentally on syntactic analysis of the questions and the Spanish documents (EFE collection), where the system tries to localize the answer. In order to make the syntactic analysis, SUPAR system is used, which works in the output of a PoS tagger.

In order to correct some of the wrong output produced by the PoS tagger and SUPAR system, a new set of post-processing rules are introduced in the system. For example, in the case of the Pos Tagger when the word "*El*" is labeled as *proper name*, a rule replaces this tag by the tag *determiner*.

- Post- processing rule: *El* NP00000 → *El* DA0MS0

SUPAR performs partial syntactic analysis that lets us to identify the syntactic blocks (SB) of the sentence which are our basic syntactic unit to define patterns. For this participation at CLEF-2006, SUPAR has allowed the nominal phrases coordinated in the appositions. These SB are classified in three types: verb phrase (VP), simple nominal phrase (NP) and simple prepositional phrase (PP).

For example:

- **Sentence:** The UEFA will present the European candidacy in Tunisia.
- **SBs:** [NP, UEFA] [VP, to present] [NP, European candidacy [PP, in: Tunisia]]

The overall architecture of our system (see figure 1) is divided in two main phases: Indexation phase and Search phase. The indexation phase consists of arranging the data where the system tries to find the answer of the questions. For example, in the QA indexation the NP, VP and PP obtained from the parsing are stored, and it also stores the results of the WSD process. On the
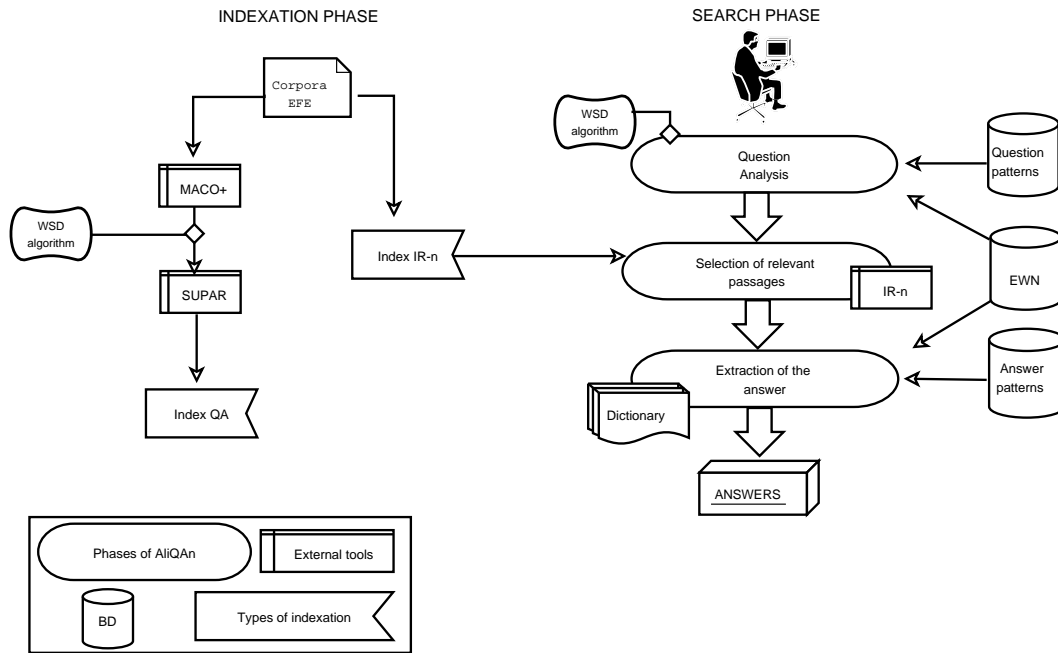

Figure 1: AliQAn system architecture

other hand, the search phase follows the most commonly used schema where three main modules are processed (Question analysis, Selection of relevant passages and Extraction of the answer). The different main characteristic in this schema is the use of dictionaries in the phase of extraction of the answer.

The used annotation is: symbols "[ ]" delimit a SB (NP, VP and PP), *"sp"* is a preposition of a PP, the term *"ap"* indicates that PP is an apposition of the previous nominal head, *SOL* is the place where the answer can be found and the symbols *"[. . .]"* indicate some irrelevant SB for the search.

2.2 Question Analysis

In this phase of the AliQAn system a new subset of syntactic patterns to detect the expected answer type is introduced, with respect to the patterns used in CLEF-2005. Besides, our taxonomy has been increased in new four types (profession, first name, temporary ephemeride and temporary day).

AliQAn carries out two main task in its step:

- To detect the type of information that the answer has to satisfy to be a candidate of answer (proper name, quantity, date . . .).

- To select the question terms (keywords) that make possible to locate those documents that can contain the answer.

The taxonomy is based on WordNet Based-Types and EuroWordNet (EWN) Top-Concepts and it is composed by the next categories: person, profession, first name, group, object, place, place city, place capital, place country, abbreviation, event, numerical quantity, numerical economic, numerical age, numerical measure, numerical period, numerical percentage, temporary year, temporary month, temporary date, temporary ephemeride, temporary day and definition.

A set of around 200 syntactic patterns are processed for the determination of the different semantic category of our ontology. The system compares the SB of the patterns to the SB of the question, the result of the comparison determines the category of the question.

The next example shows the behavior of question analysis:

- Question 06 at CLEF 2006:
    - *¿A qué país invadió Irak en 1990?* (Which country did Iraq invade in 1990?)
- Syntactic Blocks:
    - [preposition *A*]
    - [IP *qué*]
    - [NP *país*]
    - [VP *invadir*]
    - [NP *Irak* [PP, *en: 1990*]]

We have a pronoun or interrogative particle *qué (who)* followed by a syntactic blocks that contains the word *country*. This example matches with the next pattern:

[preposition *A*] [IP, *qué* | cuál] [NP, *sinónimo país* (synonymic country)]

therefore, the category of the question is *country*.

For each SB of the pattern, we keep a flag in order to determine whether the SB of the question is considered for the next stage of the QA process or not.

2.3   Selection of Relevant Passages

This module of the QA process retrieves passages using IR-n system [5]. IR-n returns a list of passages where we apply the extraction of the answer process. Besides, the objective of this task is reducing complexity of the process of searching the solution by means of reducing the amount of text in which the system searches for the answer.

The inputs of the IR-n system are changed with respect to the inputs used in CLEF-2005. For this participation at CLEF-2006 the inputs introduced to passages retrieval system are only the detected keywords in question analysis.

The next example shows the behavior of this module:

- Question 193 at CLEF 2006:
    - *¿Qué país invadió Greater Hanish?* (What country invaded the Greater Hanish?)
- Syntactic Blocks:
    - [IP *qué*]
    - [VP *invadir*]
    - [NP *Gran Hanish*]
- Type: place_country

In this question the list of words is: "invadió Greater Hanish". The word "país" has been removed because we did an analysis of the answers and we realized that in this type of questions, *place_country*, the text which contained the correct answer did not have the word "país". The same process has been applied to the questions of type *place_capital*.

2.4   Extraction of the Answer

Once the relevance passage has been retrieved, the module of extraction of the answer tries to extract a concise answer to the question.

For our CLEF-2006 participation we enhanced the type of question, SB of the question and a set of syntactic patterns with lexical, syntactic and semantic information are used in order to find a possible answer.

As shown in the next list, the system uses the following NLP techniques improved.

- *Lexical level.* Grammatical category of answer must be checked in agreement to the type of the question. For example, if we are searching for a *person*, the proposed SB as possible answer has to contain at least a noun.

- *Syntactic level.* Syntactic patterns have been redefined. Those let us to look for the answer inside the recovered passages.

- *Semantic level.* Semantic restrictions must be checked. For example, if the type of the question is *city* the possible answer must contain a hyponym of *city* in EWN. Semantic restrictions are applied in agreement to the type of the questions. Some types are not associated with semantic restrictions, such as *quantity*.

In order to design and group the patterns in several sets, the cases of the question are used. The patterns are classified in the followings three cases:

- **Case 1**. In the question, one SB of type NP or PP is only detected.

- **Case 2**. A VP is detected in the question. This verb expresses an action that must be used in order to search the answer.

- **Case 3**. VP is preceded by a NP or PP. In this case we used three sections to find out the possible answer.

According to the question and its case, the patterns are applied. In this CLEF-2006 participation we have modified the patterns. Some of them have been added and other have been subtracted. This modified has enhanced the final system result. The patterns applied to each case has been thoroughly studied. Thus, we have used 14 patterns in all.

2.4.1 Value of the Solution.

In order to select the answer from a set of candidates, each possible answer is scored.

The score of a candidate is structured in three phases: comparison of the terms inside a nominal head of a SB with the terms of the nominal head of another SB, comparison of a SB of the question with a SB of the text and weighting of a pattern according to the different SB.

- Comparison of the Terms of a Nominal Head (more details in [6]).

- Comparison of the SB. In our approach, the comparison of the SB occurs in two kinds of circumstances. When the SB of the question is localized in the text in order to apply a pattern and when the system is analizing a SB to find the answer.

  In [6] is presented the first type of comparison is called "*value of terms*", second type of comparison of SB is the calculation of the *value of solution*. It takes into account a set of evaluation rules according to the type of the question, such as:

  Lexical restrictions: grammatical category of the answer depends on the type of the question. For example, a question of type "persona (person)" the answer must have at least a proper noun or common noun. This restrictions have been refined.

  Semantic restrictions: the system leaks the answer according to semantic relations such as hyponimy. In this new AliQAn version, an new restrictions set have been added to each type of question. Dictionaries have been creaded for the countries and cities case. To that effect, AliQAn incorporate the use of this dictionaries as measure for the calculation of this restrictions. If the type of the question is *country*, the possible answer must must to belong to country dictionary. The case of the *city* type is similar to country type. For other hands, the type of questions is *capital*, as there is no such dictionary, the type is checked with the city dictionary, in this case the rank is consequently less that *city*.

  Ad-hoc restrictions: an example of this kind of restrictions is founded in the questions of type "fecha (date)", when the system penalizes the value of solution if the answer does not contain day, month and year.

- *Comparison of the Patterns and Final Evaluation of Patterns.* Theses process do not have been modified with respect to CLEF-2005

3 Cross–lingual Task

In the case of the Cross-lingual task, the BRILI [3] system (Spanish acronym for "Question Answering using Inter Lingual Index Module") is used. The fundamental characteristic of BRILI is the strategy used for the question processing module in which the Inter Lingual Index (ILI) Module of EWN is used with the aim of reducing the negative effect of question translation on the overall accuracy.

BRILI is designed to localize answers from documents, where both answers and documents are written in different languages. The system is based on complex pattern matching using NLP tools [1, 2, 5, 7]. Besides, WSD is applied to improve the precision of the system.

For the first implementation of the BRILI system, we have used the indexation phase of documents and the answer extraction module of our monolingual Spanish QA system (AliQAn) being able to answer English questions from Spanish documents.

The BRILI system introduces two improvements that alleviate the negative effect produced by the machine translation: (1) BRILI considers more than only one translation per word by means of using the different synsets of each word in the ILI module of EWN; (2) unlike the current bilingual English–Spanish QA systems, the question analysis is developed in the original language without any translation.

In order to show the complete process of BRILI, a example of a question at CLEF 2004 is detailed:

- **Question 101 at CLEF 2004:** *What army occupied Haiti?*
- **Syntactic Blocks:** [NP army]

  [VP to occupy]

  [NP Haiti]
- **Type:** group
- **Keywords to be referenced:** army occupy Haiti

  - army ↦ ejército
  - occupy ↦ absorber **ocupar** atraer residir vivir colmar rellenar **ocupar** llenar
  - Haiti ↦ Haití

- **Syntactic Blocks to search in Spanish documents:** [NP ejrcito]

  [VP ocupar]

  [NP Hait]

As it is described in the previous example, BRILI carries out a question analysis process that is based fundamentally on syntactic analysis of the question. In order to make this task, SUPAR system is used, which works in the output of a PoS tagger [7]. Besides, BRILI detects the type of information that the answer has to satisfy to be a candidate of an answer (proper name, quantity, date, ...) using a set of syntactic patterns.

Furthermore, in order to link the two languages involved in the cross-lingual QA process, the ILI module of EWN is used. In the previous example, the system finds more than one Spanish equivalents for one English word. The current strategy employed to solve this handicap consists of assigning more value to the word with the highest frequency. In the case of the previous example, the most valuated Spanish word would be *"ocupar"*.

On the other hand, the words that are not in EWN are translated into the rest of the languages using an on-line Spanish Dictionary[3]. Besides, BRILI uses gazetteers of organizations and places in order to translate words that have not linked using ILI. Therefore, in order to decrease the

---

[3]http://www.wordreference.com

Table 1: Results of AliQAn (monolingual) and BRILI(bilingual) runs.

	Right	Wrong	Inexact	Unsupported	Overall Accuracy (%)
alia061eses	72	95	15	8	37.89
alia061enes	41	124	9	16	21.58

effect of incorrect translation of the proper names, the matches using these words in the search of the answer are realized using the translated word and the original word of the question. The found matches using the original English word are valuated a 20% less.

4 Results

This section describes the results obtained in our CLEF-2006 participation. We submitted two runs. The first run was obtained for monolingual task (alia061eses) and the second is for bilingual task (alia061enes). Table 1 shows the results for each run and how theses errors lowered our system performance giving wrong answers.

CLEF-2006 competition has incorporated the list questions (5% of questions), with respect to it, our system does not consider this question type.

Unsupported answers have been increased according to evaluation for this year. The submitted snippet is more smaller that the passage obtained for our systems. Thus, our main error in this case was where to consider the begin of the submitted snippet in the passage retrieved. Two answers are printing errors for AliQAn and four answers for BRILI, the remains of the unsupported answers the passage were right.

5 Conclusion and Future Work

This paper summarizes our participation in the CLEF-2006 monolingual and bilingual QA task. This year, we have presented two system:

- Spanish monolingual QA system AliQAn. System designed to search Spanish documents in response to Spanish queries.

- Bilingual QA system BRILI. It takes questions formulated in English and the answer is searched in Spanish documents. BRILI is an extension of the AliQAn system.

The overall accuracy levels are 37.89% for AliQAn and 21.58% for BRILI. The first system has obtained a improvement of 14.81% with respect to our participation in CLEF-2005.

The AliQAn system for CLEF-2006 has been enhanced with respect to CLEF-2005.

- Lexical, morphological and syntactic errors have been detected.

- The IR engine has been modified. It only considers the words for the question analysis module as well as we do not use the question expansion of IR-n.

- A new ontology has been obtained by incorporating new elements: profession, first name, temporary ephemeride and temporary day. This modification has been needed for to add new restrictions more specific for each ontology element.

- A thorough analysis was achieved in the extraction of the answer phase. We have modified the patterns used in this phase. We have been added and removed patterns.

- New dictionaries have been used in the type of answer evaluation. These dictionaries are for the questions of type *place*.

- A New process has been used for to cut away the final answer of the questions. Thus, the inexact answers has been enhance (9 questions).

For our first participation with the BRILI system, we proposed a system that use the indexation phase of documents and the answer extraction module of AliQAn being able to answer English questions from Spanish documents. Furthermore, the two improvements above affect to this system.

Our cross-lingual system does not achieve a machine translation of the question, it indexes the words using the ILI of EWN reducing the negative effect of question translation on the overall accuracy. Moreover, the effect of incorrect translation of the proper names is decreased by using the translated word and the original word of the question.

Ongoing work on the system is focused on improve the monolingual and bilingual tasks precision, temporal question treatment and the incorporation of knowledge to those phases that can be useful to increase the our system performance.

References

[1] S. Acebo, A. Ageno, S. Climent, J. Farreres, L. Padró, R. Placer, H. Rodriguez, M. Taulé, and J. Turno. MACO: Morphological Analyzer Corpus-Oriented. *ESPRIT BRA-7315 Aquilex II, Working Paper 31*, 1994.

[2] A. Ferrández, M. Palomar, and L. Moreno. An Empirical Approach to Spanish Anaphora Resolution. *Machine Translation. Special Issue on Anaphora Resolution In Machine Translation*, 14(3/4):191–216, December 1999.

[3] S. Ferrández and A. Ferrández. Cross-lingual question answering using inter lingual index module of eurowordnet. *Advances in Natural Language Processing. Research in Computing Science. ISSN: 1665-9899*, 18:177–182, February 2006.

[4] S. Ferrández, S. Roger, A. Ferrández, A. Aguilar, and P. López-Moreno. A new proposal of word sense disambiguation for nouns on a question answering system. *Advances in Natural Language Processing. Research in Computing Science. ISSN: 1665-9899*, 18:83–92, February 2006.

[5] F. Llopis and J.L. Vicedo. Ir-n, a passage retrieval system. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, 2001.

[6] S. Roger, S. Ferrández, A. Ferrández, J. Peral, F. Llopis, A. Aguilar, and D. Tomás. AliQAn, Spanish QA System at CLEF-2005. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, 2005.

[7] H. Schmid. TreeTagger — a language independent part-of-speech tagger. *Institut fur Maschinelle Sprachverarbeitung, Universitat Stuttgart*, 1995.