AliQAn, Spanish QA System at CLEF-2008*

S. Roger, K. Vila, A. Ferrández, M. Pardiño, J. M. Gómez, M. Puchol-Blasco and J. Peral.
Natural Language Processing and Information Systems Group.
Department of Software and Computing Systems.
University of Alicante, Spain.
{sroger,kvila,antonio,maria,jmgomez,marcel,jperal}@dlsi.ua.es

Abstract

This paper describes the participation of the system AliQAn, a monolingual open-domain Question Answering (QA) System developed in the Department of Language Processing and Information System at the University of Alicante, in the CLEF-2008 Spanish monolingual QA evaluation task. Here, we focus on explaining a couple of strong points of the current version of AliQAn: (i) our algorithm for dealing with topic-related questions, and (ii) our approach for decreasing the number of inexact answers. We have also explored the use of the Wikipedia corpora, which have proposed some new challenges for the QA task. Besides, the achieved results (overall accuracy of 19.50%) are shown and discussed 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

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids

1 Introduction

Our system has been developed during three participations (2005, 2006 and 2008) in the Department of Language Processing and Information Systems at the University of Alicante. It is based

*This work has been partially supported by the framework of the project QALL-ME (FP6-IST-033860), which is a 6th Framenwork Research Programme of the European Union (EU), the Spanish Government, project TEXT-MESS (TIN-2006-15265-C06-01), by the University of Comahue under the project 04/E062, by the Generalitat Valenciana throught the research grants BFPI/2008/093 and BFPI06/182 and University of Matanzas.

on complex pattern matching using NLP tools. This year we have adapted our system to work on Wikipedia and we have proposed a method to work with inexact answers. None method for anaphora resolution has been done for the topic-related questions. A simple method has been proposed to treat this type of questions.

The rest of this paper is organized as follows: section two and three describes our algorithm for dealing with topic-related questions and the special treatment of inexact answers respectively. Afterwards, the handling of Wikipedia and related problems. Section five describes the obtained results. Finally, our conclusions and future work are presented.

2 Dealing with Topic-Related Questions

Before QA@CLEF 2007, all questions could be answered in isolation without any reference to the context (previous questions or answers). From QA@CLEF 2007 run, topic-related questions are clusters of questions which are related to the same topic and possibly contain anaphoric references between one question and the other questions of the same cluster.

For example, the questions of QA@CLEF 2008: 008-"¿Dónde vivía la tribu de los Mojave?" (Where did the Mohave tribe live?) and 009: "¿Quiénes eran sus enemigos?" (Who were their enemies?) belong to the same cluster. On the other hand, the following example shows an anaphoric reference between second question and third question of the one cluster: 029: "¿Entre qué días fue la batalla de Brunete?" (Among which days was the battle of Brunete?), 030: "¿Dónde se publicó el reportaje de Gerda Taro sobre esta batalla?" (Where was the article of Gerda Taro about this battle published?) and 031: ¿A qué hospital fue trasladada tras su accidente? (Which hospital were she moved to after her accident?). In 2007, 30 questions out of 200 were topic-related and in this year this size was enlarged to 64 out of 200 for Spanish.

To treat such context-dependent questions, the underlying idea was very simple. It considers the enrichment of dependent questions by adding some noun phrases of the first question of each cluster and the noun of the answer for this question. By reasons of simplicity and for avoid introducing noise; we only considered the co-reference between the first question and other of the same cluster. The algorithm employed contained the following steps:

1. Answering the first question of one cluster without special treatment.
2. Extracting the set of noun phrases from question and answer.
3. Adding this set of noun phrases to all dependent questions.
4. Handling and extracting the answers from these expanded questions.

For instance, if we consider the question 008, the system returns the answer "Arizona"(step 1). Step 2 produces the noun phrases "la tribu de los Mojave" (Mohave tribe) and "Arizona". Then (step 3), it obtains the noun phrases that correspond to the 009 question: "sus enemigos" (their enemies). Finally, in step 4, the previous extended set with all the noun phrases is used to find the answer of question 009.

For the final answers to topic-related questions we obtain the following criteria:

- If the answer to the extended question (following the steps previously described) was *nil*, then *nil* was returned as final answer.

- In the opposite case, we have ranked the answers (with or without extension) in a decreasing order, thus returning the first three ones of the ranking.

3 Inexact Treatment

This year, the algorithm to treat the inexact answers has been modified only for the questions which expected answer type as group, person, first name, place, country and city. In theses cases, we considered that the answers are able to contain one or more noun phrases. Before explaining the general algorithm used for handling theses answers, we will define some variables: Set $NP = np_1, np_2, \ldots, np_n$ where np_i $\forall i = 1..n$ are consecutive proper names of the head of the different noun phrases (included embebed noun phrases). Set $\Omega = \{NP | NP \in$ head of the noun phrases of the answer$\}$. For example, let considere the sentence "la Sociedad Española de Vexilología" (the Spanish Society of Vexillology), in this case the cardinality of Ω ($|\Omega|$) is 2 and its elements are: "Spanish Society" and "Vexillology".

The algorithm begins by finding the set Ω for all noun phrases of the answer. If $|\Omega| > 1$ then the elements of Ω are ranking according to its weight. The weight is increased or decreased in accordance with many different criteria and whether it belongs to specific dictionary or it does not. Criteria and dictionary are defined according to expected answer type. After this, the algorithm selects the element with bigger score and it returns the head of the noun phrase corresponding to this element. On the other hand, if the $|\Omega| = 1$, then it only returns the corresponding head of the noun phrase. In the previous example, we suppose that the weight of "Spanish Society" is N_1 and the weight of "Vexillology" is N_2. If $N_1 > N_2$, then algorithm returns "the Spanish Society" else it returns "Vexillology".

4 Exploring Wikipedia

Compared to traditional CLEF corpora (based on articles from newspapers), Wikipedia is a very large document collection and has not enough redundancy. In spite of that fact, the articles from newspapers have a fair amount of redundancy because they are usually published, with pretty much relevance, in different days, by different people and using different expressions. Wikipedia collections use hyperlinks to avoid information repetition (i.e. data which is sensitive to be repeated is replaced by links to the original source).

An Information Retrieval (IR) system needs to be more precise in order to filter the fair amount of irrelevant information due to the size of the Wikipedia collections. At the same time, an IR system needs to have high coverage to deal with the low redundancy of these corpora. In addition, Wikipedia, unlike newspaper collections, is highly structured. This structure gives a lot of information about the article topic in the form of tables, references and links. Hence, an IR system needs to consider this structure to take advantage of this information.

Bearing these considerations in mind, we aim to adapt two IR systems (namely, IR-n [5] and JIRS [4]) in order to (i) be able to use very large document collections, and (ii) face up to the above-commented new Wikipedia challenges. Specifically, in this paper, our effort has been directed towards solving the first goal.

In addition, we would like to point out that several problems derived from the codification of the no-latin characters in Wikipedia were solved from the viewpoint of our QA system. The

Table 1: General results obtained in the QA@CLEF-2008

Right	39
Inexact	4
Unsupported	1
Accuracy (overall)	19.50%
Factoid questions	20.49 %
Definition questions	31.57 %
List questions	0%
Temporally questions	14.28 %

source of these problems is that the Wikipedia collections was coded in UTF-8, while our QA system uses ISO encoding to perform the morpho-syntactic labelling of documents via MACO [1] and SUPAR [2] NLP tools.

An example that illustrates this problem is shown in the following question from QA@CLEF 2008: 048- "¿Qué cargo ocupaba Hideki Tōjō antes del ataque a Pearl Harbor?" (What position did Hideki Tōjō hold before the Pearl Harbor attack?) the "ō" character was codified as "?" by our system.

Our proposed solution for our QA system consists of controlling the correspondences between the two encodings for non-latin characters. Even though it is a very simple solution, good results are obtained. Nevertheless, as future work we wish to adapt our system and its related tools to directly work the UTF-8 encoding.

5 Results

This section describes the table related with the results and the evaluation of our system in CLEF-2008. The proposed system was applied to the set of 200 questions.

The University submitted four runs to QA@CLEF, two runs for Spanish-Spanish (spsp) and two runs for English-Spanish (ensp). We participate with the monolingual runs. Regrettably, there was an error in the submitting of the runs, one run of spsp was overwritten with one run of ensp. Therefore, we only look the good run.

Table 1 shows the results for this run and the effect that the errors and problems produced our system performance. It is important to remark that it is our first participation with the new characteristic: Wikipedia and topic-related questions.

On the other hand, AliQAn system had a high percentage of inexact answers in previous years. This kind of answers has been improved in this participation: of 24 in the year 2005 [6] and 15 in the year 2006 [3] to 4 this year (2008), which all correspond a list questions[1].

6 Conclusion and Future Work

This paper summarizes our participation in the CLEF-2008 monolingual task with our monolingual open-domain QA System (AliQAn). This year, the main contributions were:

[1] It is important to say that list questions are not supported by our system.

- Algorithm for resolving topic-related questions. The essence of this algorithm is to extend every question (q_i) by adding some noun phrases and the noun of the answer of the first question of the same cluster which q_i depends on.
- Approach for decreasing the number of inexact answers. This approach assigns certain weight (determined by using specific dictionaries) to the heads of each answer's noun phrase according to an expected answer type and it returns the head of the noun phrase with the greatest weight. We have obtained excellent results with a decrease of 20 inexact answers with regard to the year 2005.
- Using Wikipedia with our IR & QA systems. On one hand, our IR system has been adapted for making possible the use Wikipedia with very large document collections. On the other hand, several problems derived from the codification of the non-latin characters in Wikipedia have been resolved in order to properly use it together with our QA system.

All questions given in this track, except the list questions, have been treated by our system and only one has been unsupported. Our paper only includes one run for the Spanish monolingual QA task and it has achieved an overall accuracy of 19.50%. Finally, we would like to point out that this is the first time we deal with Wikipedia and topic-related questions for our participation in the CLEF QA task.

Our future work is focused on the multilingual task, the adaptation of the NLP tools related to our system to directly work the UTF-8 encoding and the incorporation of knowledge to the phases that can be useful to increase the performance of our system.

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, P. López-Moreno, S. Roger, A. Ferrández, J. Peral, X. Alvarado, E. Noguera, and F. Llopis. Monolingual and cross-lingual qa using aliqan and brili systems for clef 2006. In *CLEF*, pages 450–453, 2006.

[4] J. M. Gómez, M. Montes-Gómez, E. Sanchis, L. Villaseor-Pineda, and P. Rosso. Language independent passage retrieval for question answering. In *Fourth Mexican International Conference on Artificial Intelligence MICAI 2005*, Lecture Notes in Computer Science, pages 816–823, Monterrey, Mexico, 2005. Springer-Verlag GmbH.

[5] F. Llopis, J. L. Vicedo, and A. Ferrández. Passage selection to improve question answering. In *In Proceedings of the COLING 2002 Workshop on Multilingual Summarization and Question Answering*, pages 1–6, Taipei, Taiwan, 2002.

[6] S. Roger, S. Ferrández, A. Ferrández, J. Peral, F. Llopis, A. Aguilar, and D. Tomás. Aliqan, spanish qa system at clef-2005. In *CLEF*, pages 457–466, 2005.