AliQAn, Spanish QA System at multilingual QA@CLEF-2008*

R. Muñoz-Terol, M.Puchol-Blasco, M. Pardiño, J.M. Gómez, S.Roger, K. Vila,
A. Ferrández, J. Peral, P. Martínez-Barco
Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Natural Language Processing and Information Systems Group Department of Software
and Computing Systems
University of Alicante, Spain
rafamt,marcel,maria,jmgomez,sroger,kvila,antonio,jperal,patricio@dlsi.ua.es

Abstract

In QA@CLEF 2008, we participate in monolingual (Spanish) and multilingual (English - Spanish) tasks. Specifically, in this paper, we will tackle with the English - Spanish QA task. In this edition we will deal with two main problems: an heterogeneous document collection (news articles and Wikipedia) and a large number of topic-related questions, which make somewhat difficult our participation. We want to highlight in the translation module in our system two possible mechanisms: one based on logic forms, and the other, on machine translation techniques. In addition, it has also been used a system of anaphora resolution that it is described below and a QA System, AliQAn (also used this year in the monolingual task).

1 Introduction

For our participation in this task we have used a monolingual QA system: AliQAn; a translation system based on logic forms and an anaphora resolution system to address the interrelated questions and answers.
The system AliQAn was widely described by Roger *et al.* [8] after their participation in the competition of the Spanish CLEF 2005. It is important to emphasize that this system has been adapted to work in this year's task monolingual too, since there are a great number of changes made in Roger *et al.* [9].
The main problems that contributed to make difficult our participation were the processing of Wikipedia (this corpus is different from those used to work with AliQAn) and the large number of questions related.

In this research work, our QA system applies two different methods to perform the translation of the questions from one language to another. The first one consists in applying natural language processing techniques based on the formal representation of questions, by using logic forms [10] and, the second one, simply consists in using machine translation techniques. So, the main goal consists in comparing both methods in the framework of the efficiency of the question answering system to solve the question answering problem.

The anaphora resolution problem has been introduced in the last two years in QA task, increasing the number of questions with anaphoric relationships since last year. As a consequence, a

*This paper has been partially supported by the Spanish government, project TIN-2006-15265-C06-01 and project GV06/028, and by the framework of the project QALL-ME, which is a 6th Framework Research Programme of the European Union (EU), contract number: FP6-IST-033860, by the Generalitat Valenciana throught the research grants BFPI/2008/093 and BFPI06/182 and by the University of Comahue under the project 04/E062.

spanish rule-based anaphora resolution system for QA has been implemented to solve this problem. In our system, three kinds of anaphora can be solved: pronoun anaphora, definite descriptions, and zero anaphora. The results obtained by our anaphora resolution system are around 40% precision, using the questions translated from English into Spanish.

2 Background

Previous research works such as Bowden *et al.* [1] directly apply machine translation methods to translate the questions. On the contrary, and in order to perform the translation of the questions from one language to other one, the research work developed by Ferrández and Ferrández [4] and the PhD thesis defended by Ferrández [3] exhibits that the use of machine translation techniques to perform the translation of questions from one language to another decreases the efficiency of the multilingual question answering system versus other approaches based on natural language processing techniques.

3 Description of the System

3.1 English-Spanish Translation based on Logic Forms and lexical resources

The translation of questions from English into Spanish is performed by inferring the logic form of the questions and using lexical resources to translate the predicates of the logic form. The technique applied to infer the logic form of the questions is the one developed by Terol *et al.* [10]. Consequently, this translation technique is performed as follows:

- The predicates whose types are corresponded to noun or verb are translated using the EuroWordNet [11] lexical resource. The connection between the synsets of the English and Spanish WordNets is performed in a similar way as treated by Ferrandez *et al.* [5] using the Inter-Lingual-Index (ILI). Each one of the synsets that are mapped from English into Spanish towards ILI contain a set of concepts. The processing consists in counting the occurrences of the different concepts that appear in the mapping process and, finally, the concept with the highest number of occurrences is chosen as the translation of the predicate.

- The predicates whose type is instantiated to adjective, adverb or preposition, and the predicates treated in the previous step that are not translated by EuroWordNet are translated applying the English-Spanish Babylon dictionary [1]. As it occurred in the previous step, the dictionary can return a set of different translations grouped in synsets (different from WordNet synsets). Thus, the processing consists in counting the occurrences of the different translations that the dictionary returns and, finally, the translation with the highest number of occurrences is chosen as the translation of the predicate.

- Finally, the rest of predicates and the ones than are not translated in the two previous steps are definitely translated by using the Google Translation Toolkit [2].

Once the predicates of the logic form are translated according to the recently detailed rules, the last translation task consists in translating the question as a result of applying some English-Spanish contrastive grammar rules to the sequence of translations of predicates of the logic form. The applied English-Spanish contrastive grammar rules are based on the ones derived from the

[1] http://www.babylon.com
[2] http://www.google.com/translate_t

previous study developed by Fernandez *et al.* [2] and Martinez-Vazquez [6]. The applied English-Spanish contrastive grammar rules are detailed in Table 1, and an example of translation, applying each one of these rules, is explained in Table 2. Note that the rule number four is recurrent because a complex nominal (NNC) can recurringly derive others.

Rule Id.	English Structure	Spanish Translation
1	JJ + NN	*Translation(NN)* + *Translation(JJ)*
2	JJ_1 + JJ_2 + NN	*Translation*(JJ_1) + *Translation(NN)* + *Translation*(JJ_2)
3	NN_1 + NN_2	*Translation*(NN_1) + *Translation*(NN_2)
4	NN + NNC	*Translation(NNC)* + "de" + *Translation(NN)*
5	JJ + NN_1 + NN_2	*Translation*(NN_2) + *Translation(JJ)* + "de" + *Translation*(NN_1)

Table 1: Applied English-Spanish contrastive grammar rules

Rule Id.	English Expression	Spanish Translation
1	*red car*	*coche rojo*
2	*beautiful green eyes*	*bonitos ojos verdes*
3	*electronic database*	*base de datos electrnica*
4	*train station ticket office*	*oficina de billete de estacin de tren*
5	*multiple regression model*	*modelo mltiple de regresin*

Table 2: Example of the applied English-Spanish contrastive grammar rules

Finally, the translation performed by the rest of predicates of the logic form, whose logic structure does not match these English-Spanish contrastive grammar rules, consists in the concatenation of the sequence of translations of these predicates.

3.2 Anaphora Resolution

As last year, anaphora resolution problem has been introduced this year. Questions are grouped in topics. In those topics, the first question is anaphora-free, but the other questions may require information from data contained in the first question, or in the first answer *(it can only be the first question or the first answer for a specific topic, exactly as described in the guidelines for this year)*.

From our analysis of Spanish examples used last year, we have discovered three types of possible anaphora: pronoun anaphora, definite descriptions, and zero anaphora. These types of anaphora have been analyzed and are the basis for our system. Our approach for anaphora resolution is rule-based. Basically, it is based on the papers referenced in Mitkov [7].

The first step is to get the Part-of-Speech tagging for possible antecedents *(the first question or the first answer for an specific topic)* and a question with a possible anaphora. Later, an anaphora detection module is executed. This module can be considered as three different subsystems in which each subsystem corrsponds to: pronoun anaphora detection, definite description detection and zero anaphora detection. In the first one, pronouns are taken as anaphora. In definite descriptions, noun phrases are taken as a possible anaphora. And, in zero anaphora, the entire phrase can be taken as a possible anaphora.

For all anaphora resolution modules, for each topic group, noun phrases are extracted from the first question-answer pair as possible antecedent. Then, when we have a pronoun anaphora, the possible anaphora is compared in gender and number with each possible antecedent. In definite descriptions and zero anaphora, for each noun phrase, a google search is launched joining the possible antecedent and the main words of the anaphora *(or main words contained in the noun phrases, in case of zero anaphora)*. Later, for each noun phrase in the possible anaphora,

the relations between it and the main words contained in the antecedents are extracted using MultiWordNet [3] .

In all previous cases, a specific weight is assigned to each possible anaphora resolution. Later, those weights are ordered, and the best case is selected as anaphora resolution for the related question. It is important to mention that if the noun phrase is in the answer, the greater weight is assigned to it, due to the analysis done in the last year corpus.

A few manual experiments are performed using the last year Spanish corpus. The results obtained are around 56% precision. The results obtained using the corpus for this year, once the questions are translated from English into Spanish, are worse, obtaining around 40% precision.

4 Results and Conclusions

The scores obtained in the application of the machine translation techniques are a bit better than the ones obtained in the application of the techniques based on logic forms. This can be due to the fact that the use of logic forms is a good method to perform the language-independent knowledge representation, but this method must be improved to perform the translation of sentences from one language to another. As a further work, the next research goal will be to improve the logic form processing methods in the translation process.

We have had some problems with the anaphora resolution system, because it is a rule-based system and it needs well-formed sentences to work. Due to the fact that translations offered by the system are not well-formed, in most cases, the precision obtained by the anaphora resolution system has decreased considerably, arriving at around 40% precision.

Finally, we present the results obtained with our system in this task in Tables 3 and 4. Specifically, Tables 3 shows the number of right answers(R), wrong answers (W), inexact answers (X), unsupported answers (U) obtained by our system.

Run_ID	**#Questions**	**#R**	**#W**	**#X**	**#U**
Run 1 (machine translation)	200	25	173	0	2
Run 2 (logic forms)	200	18	176	3	3

Table 3: Results obtained by our system at English-Spanish QA task 2008

Table 4 shows the comparative of the values obtained for MRR, CSW and Accuracy with both runs.

Run_ID	**# Questions**	**Accuracy(%)**	**CSW**	**MRR**
Run 1 (machine translation)	200	12.5	0.01114	0.17797
Run 2 (logic forms)	200	9.0	0.00626	0.11499

Table 4: Results obtained by our system at English-Spanish QA task 2008

References

[1] Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Thomas d'Silva, and Dan Moldovan. Multilingual question answering through intermediate translation: Lcc's poweranswer at qa@clef 2007. In *Working Notes for the CLEF 2007 Workshop*, 2007.

[2] F. Fernández and B. Montero-Fleta. *La premodificación nominal en el ámbito de la informtica. Estudio contrastivo inglés-español.* Universidad de Valencia, 2003.

[3] `http://multiwordnet.itc.it/online/multiwordnet.php`

[3] S. Ferrández. *Arquitectura multilingüe de sistemas de Búsqueda de Respuestas basada en ILI y Wikipedia*. PhD thesis, University of Alicante, 2008.

[4] S. Ferrández and A. Ferrández. The negative effect of machine translation on crosslingual question answering. In *Computational Linguistics and Intelligent Text Processing*, pages 494–505, 2007.

[5] S. Ferrández, A. Ferrández, S. Roger, P. López-Moreno, and J. Peral. Brili, an english-spanish question answering system. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 23–29, 2006.

[6] M. Martínez-Vázquez. *Gramática contrastiva inglés-español*. Servicio de publicaciones de la Universidad de Huelva, 1996.

[7] Ruslan Mitkov. *Anaphora Resolution*. Longman, London, 2002.

[8] 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.

[9] S. Roger, K. Vila, A. Ferrández, M. Pardiño, J.M. Gómez, M. Puchol-Blasco, and J. Peral. Aliqan, spanish qa system at clef-2008. In *CLEF*, 2008.

[10] R.M. Terol, P. Martínez-Barco, and M. Palomar. A knowledge based method for the medical question answering problem. In *Computers in Biology and Medicine*, volume 37, pages 1511 – 1521, 2007.

[11] P. Vossen. *EuroWordNet General Document. Part A. Final Document.* EuroWordNet (LE2-4003, LE4-8328), 2002.