Cross-Lingual Question Answering Using Inter Lingual Index Module of EuroWordNet*

Sergio Ferrández and Antonio Ferrández

Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{sferrandez,antonio}@dlsi.ua.es

Abstract. This paper outlines the BRILI cross-lingual English-Spanish-Catalan Question Answering (QA) system. The BRILI is being designed at University of Alicante and will be capable to answer English, Spanish and Catalan questions from English, Spanish and Catalan documents. The starting point is our monolingual Spanish QA system [11] which was presented at the 2005 edition of the Cross-Language Evaluation Forum (CLEF). We describe the extensions to our monolingual QA system that are required, especially the language identification module and the strategy used for the question processing module. The Inter Lingual Index (ILI) Module of EuroWordNet (EWN) is used by the question processing modules. The aim of this is to reduce the negative effect of question translation on the overall accuracy of QA systems.

1 Introduction

The aim of a Question Answering (QA) system is to localize the correct answer to a question in natural language in a non-structured collection of documents, also the situations where the system is not to able to provide an answer should be detected. In the case of a Cross-Lingual QA (CL-QA) system, the question is formulated in a language different from the one of the documents, which increases the difficulty. Nowadays, multilingual QA systems have been recognized as an important issue for the future of Information Retrieval (IR).

In this paper we present BRILI (Spanish acronym for "Question Answering using Inter Lingual Index Module"). It is a CL-QA system for Spanish, English and Catalan. It is designed to localize answers from documents, where both answers and documents are written in the three languages. The system is based on complex pattern matching using NLP tools [1, 4, 7, 12]. Beside, Word Sense Disambiguation (WSD) is applied to improve the system (a new proposal of WSD for nouns based on [2]).

BRILI is fully automatic, including the modules of language identification and question processing. The main goal of this paper is to describe these modules

* This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia Government under project number GV04B-268.

© A. Gelbukh (Editor)
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 177-182

and the use of the ILI Module of EuroWordNet (EWN) in order to reduce the negative effects of question translation on the overall accuracy of QA systems.

The rest of this paper is organized as follows: section 2 describes effects of question translation on the precision of QA systems. Afterwards, the architecture of the system is shown and discussed in section 3 (specially the modules of language identification and question processing) and finally, section 4 details our conclusions and future work.

2 Negative effects of question translation

Nowadays, most of the implementations of CL-QA systems are based on four different approaches. The first one uses a translation system to translate question into the language in which the documents are written (this technique is the most used). On the other hand, other systems base their implementations on cross-lingual IR (CL-IR) systems. These implementations use the language of the question to generate queries to a CL-IR system. Besides, some systems translate all the documents into the language of the question. Finally, there are sophisticated implementations [6] where English is used as pivot language.

The low quality of machine translation provides results worse than those obtained in the monolingual task.

The precision of a CL-QA system is mainly affected by a correct translation and analysis of the questions that are received as input. An imperfect translation of the question causes a negative impact on the overall accuracy of the systems [3, 5, 6, 10, 13]. As Moldovan [9] stated, Question Analysis phase is responsible for 36.4% of the total of number of errors in open-domain QA.

For CL-QA, translations are often inexact and quite fuzzy, this fact causes an important decrease on the precision of the systems. For instance, on-line Machines Translation systems generate errors such as translations of names that should be left untranslated. The impact of this kind of mistakes should be controlled and valued.

The number of correct answers is always lower for CL-QA. Usually, the precision on cross-lingual task is approximately 50% lower than for monolingual task [14].

3 Architecture Overview

The starting point of BRILI is our monolingual Spanish QA system, called AliQAn [11], which was presented at the 2005 edition of the Cross-Language Evaluation Forum (CLEF).

AliQAn is based fundamentally on syntactic analysis of the questions and the Spanish documents, where the system tries to localize the answer. In order to make the syntactic analysis, SUPAR [4] system is used, which works in the output of a PoS tagger [1]. SUPAR performs partial syntactic analysis that lets us to identify the different grammatical structures of the sentence. Syntactic blocks (SB) are extracted, and they are our basic syntactic unit to define patterns.

Using the output of SUPAR, AliQAn identifies three types of SB: verb phrase (VP), simple nominal phrase (NP) and simple prepositional phrase (PP).

In the step of the extraction of the answer, AliQAn takes the set of retrieved passages by IR-n[7] and tries to extract a concise answer to the question. In order to extract the answer, the following NLP techniques are used:

- *Lexical level.* Grammatical category of answer must be checked according 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 defined. 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 EuroWordNet. Semantic restrictions are applied according to the type of the questions. Some types are not associated with semantic restrictions, such as *quantity.*

The next example (question 38, *In Workshop CLEF 2003*) shows the used pattern and the behavior the extraction of the answer:

- [SOL[PP, sp: NP1]] [...] [VP][...] [NP2]

First, NP2 (or PP2) and VP are searched by the system, afterward the NP1 with the answer must be found. Next example shows the process:

- **Question:** ¿Qué presidente de Corea del Norte murió a los 82 años de edad? (What North Korea's president died at the age of 82?)
- **Type:** person
- **List of SB:** [NP, north*korea*president] [VP, to death] [PP, at: age [PP, of: 80]
- **Text:** [...] *Kim Il Sung*, **presidente de Corea del Norte, murió ayer a los 82 años** [...] ([...] *Kim Il Sung*, **president of North Korea, died yesterday at the age of 82** [...])
- **List of SB of sentence:** [...] [NP, kim*il* sung [PP, apposition: president [PP, of: north*korea]]] [VP, to death] [PP, at: age [PP, of: 82] [...]
- **Answer:** Kim Il Sung

BRILI (the architecture is shown in Fig.1) is an automatic and complete system for CL-QA tasks. The system carries out an indexation phase where all the documents are analyzed each one in its own language, in which syntactic and semantic information is stored.

The module of language identification have been developed to automatically distinguish the correct language of the question and documents. It is based on two main techniques: the use of dictionaries (joined dictionaries, specific per-language stopwords) and the use of part-of-word terminology (for example, "ing" in the case of English). This philosophy presents a good precision [8] in Spanish, English and Catalan text.

The phase of Question Analysis is made up of two tasks, the first one consisting on detecting the type of information that the answer has to satisfy to be a candidate of answer (proper name, quantity, date, etcetera), while the second one has as objective selecting the question terms (keywords) that make possible

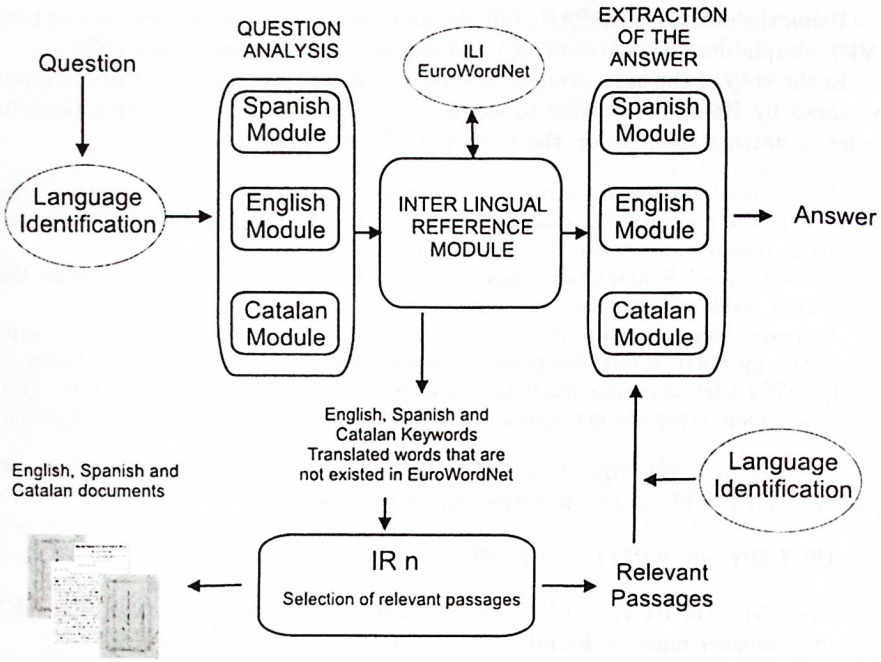

Fig. 1. System architecture

to locate those documents that can contain the answer. The expected answer type is achieved using different sets of syntactic patterns according to the language that is being processed. Beside, WSD is applied to obtain the synset for each keyword.

The next example shows the behavior of question analysis in a question of type *person*:

- Question:
  - *Which French president inaugurated the Eurotunnel?*
- Information used (syntactic blocks) to detect the type of the question and keywords:
  - interrogative pronoun - *Which*
  - nominal phrase - *French president* - synset in English
  - verb phrase - *to inaugurate* - synset in English
  - nominal phrase - *Eurotunnel* - this word does not exit in EuroWoedNet

The inputs of the Inter Lingual Reference (ILR) module are the detected keywords in the question and the type of the question. In addition to the syntactic blocks and the type of the question, the ILR module returns for each keywords its synsets in the three languages, using the ILI Module of EWN. The words that are not in EWN are translated into the rest of the languages using a machine translation system. Our approach does not achieve a translation of the

question, it indexes the words using the ILI of EWN reducing the negative effect of question translation on the overall accuracy. For instance, using the previous example the ILR module selected the synset in Spanish and Catalan for the word *"French"*, on the other hand the word *"Eurotunnel"* which is not presented in EWN is translated into Spanish and English using machine translation.

The behavior of this module is shown using the previous example.

- **Input of ILR module:**
  - *French* synset in English
  - *president* synset in English
  - *to inagurate* synset in English
  - *Eurotunnel* this word does not exit in EuroWoedNet
- **Output of ILR module:**
  - *French* synsets in English, Spanish and Catalan
  - *president* synsets in English, Spanish and Catalan
  - *to inagurate* synsets in English, Spanish and Catalan
  - *Eurotunnel* this word does not exit in EuroWoedNet
    - translations into Spanish and Catalan

The phase of Selection of relevant passages uses IR-n system [7]. The inputs of IR-n are the detected keywords and the translated word that are not in EWN. For instance, using the previous example, IR-n receives as input the words: *"inaugurate"* with its synonymous; *"inaugurar"* (in Spanish) with its synonymous and *"inaugurar"* (in Catalan) with its synonymous. The translated words have been obtained indexing the synsets of words using ILI Module of EWN without using any machine translation. IR-n returns a list of passages where the system applies the extraction of the answer according to the language in which each passage is written.

The final step of BRILI is the phase of Extraction of the Answer which is composed of three monolingual modules. BRILI uses the syntactic blocks of the question and different sets of syntactic patterns (according to the language) with lexical, syntactic and semantic information to find out the correct answer. Next, an example of syntactic pattern for Spanish is shown which captures solution in Spanish sentence.

- **Sentence:**
  "... el Presidente Francés, Jacques Chirac, inauguró el Eurotunel ..." (... the French President, Jacques Chirac, inaugurated the Eurotunnel ...)
- **Syntactic pattern:**
  [NP (*"French president"*), apposition [NP (SOLUTION)]] + [VP (*"to inaugurate"*)] + [NP *"Eurotunnel"*)]

In order to decreased the effect of incorrect translation of the words that are not in EWN, the matches using these words in the search of the answer are valued less than the words obtained from the ILI Module of EWN.

4 Conclusions and Future Work

The main objective pursued by our proposal is a cross-lingual QA system that reduces the use of machine translation. Reducing the decrease of the precision that is caused by the question translation will be achieve by means of the use of the ILI Module of EWN and WSD. Nowadays, BRILI is being implemented. For this reason, results are not presented in this paper.

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. E. Agirre and G. Rigau. A proposal for word sense disambiguation using conceptual distance. *1st Intl. Conf. on recent Advances in NLP. Bulgaria*, 1995.
3. C. de Pablo-Sánchez, A. González-Ledesma, J.L. Martínez-Fernández, J.M. Guirao, P. Martinez, and A. Moreno. MIRACLE's 2005 Approach to Cross-Lingual Question Answering. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, 2005.
4. 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.
5. J. M. Gómez, E. Bisbal, D. Buscaldi, and P. Rosso E. Sanchis. Monolingul and Cross-Language QA using a QA-oriented Passage Retrieval System. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2005.
6. D. Laurent, P. Séguéla, and S. Negre. Cross Lingual Question Answering using QRISTAL for CLEF 2005. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2005.
7. F. Llopis and J.L. Vicedo. Ir-n, a passage retrieval system. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, 2001.
8. T. Martínez, E. Noguera, R. Muñoz, and F. Llopis. Web track for CLEF2005 at ALICANTE UNIVERSITY. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2005.
9. D.I. Moldovan, M. Pasca, S.M. Harabagiu, and M. Surdeanu. Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst*, 21:133–154, 2003.
10. G. Neumann and B. Sacaleanu. DFKI's LT-lab at the CLEF 2005 Multiple Language Question Answering Track. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2005.
11. 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.
12. H. Schmid. TreeTagger — a language independent part-of-speech tagger. *Institut fur Maschinelle Sprachverarbeitung, Universitat Stuttgart*, 1995.
13. R. F. E. Sutcliffe, M. Mulcahy, and I. Gabbay. Cross-Language French-English Question Answering Using the DLT system at CLEF 2005. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2005.
14. A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutclife. Overview of the CLEF 2005 Multilingual Question Answering Track. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2005.