Cross-Language Question Answering at the USC Information Sciences Institute

Abdessamad Echihabi, Douglas W. Oard*, Daniel Marcu and Ulf Hermjakob

University of Southern California Information Sciences Institute
4676 Admiralty Way, Marina Del Rey, CA 90292 USA
echihabi@isi.edu, oard@umd.edu, marcu@isi.edu, ulf@isi.edu

Abstract. The TextMap-TMT cross-language question answering system at USC-ISI was designed to answer Spanish questions from English documents. The system is fully automatic, including question translation from Spanish to English, question type determination, rewriting to generate expected answer structures, search in the target collection and on the Web as a side collection, and answer selection from among the plausible candidates that were found. A development test collection with answer patterns for 100 questions in English and Spanish was used to assess the effect of question translation on each processing stage, and some adjustments were made to the question translation process to minimize these effects. Two runs were submitted, both of which sought to return exact answers. For the better of the two runs (which omitted an additional Web-based answer validation stage), the top-ranked answer was scored as correct in 56 of 200 cases, 53 of which were judged to be supported by the content of the target collection.

1 Introduction

The goal of a Question Answering (QA) system is to find answers to questions in a large collection of documents. The QA track in the 2003 Cross-Language Evaluation Forum (CLEF) added a new wrinkle, with the question posed in one language (Spanish, in our case) and the documents written in another (for us, English). The challenge therefore was to identify answers to a Spanish question in a collection of English documents. The focus of the evaluation was on finding answers, so translation of the answer from English into Spanish was not required.

At the University of Southern California Information Sciences Institute we have been working on question answering in English for several years. We therefore adapted our existing TextMap English QA system [3] to perform Cross-Language question answering (CL-QA) by making the following changes:

– Question translation from Spanish to English using an off-the-shelf machine translation system, augmented with some simple postprocessing to correct observed systematic errors.

* Permanent address: College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park MD 20740 USA

– Candidate generation using the Inquery text retrieval system, to capitalize on cross-language search capabilities that can be constructed more easily using Inquery than with the MG system that we had previously used.
– Answer validation, using a second Web search to increase the score of answers found in the target collection that also have support on the Web.

Imperfect translation of questions can introduce new challenges for downstream components that vary with the design of each component. To assess the impact that imperfect translation may have on the accuracy of a question answering system, we prepared a 100-question development collection consisting of questions in English, human translations of those questions into Spanish, and answer patterns that facilitated automated scoring in English. In the next section, we describe a set of experiments using this collection and explain how those experiments informed the design of our CL-QA system. Section 3 describes the architecture and main components of our system; section 4 presents our results for both one-best and three-best scoring. Finally, we draw some conclusions from this first experience with CL-QA in section 5, along with some comments about directions that we may explore in the future.

2 Measuring Question Translation Effects

There are many alternatives one can use to build a CL-QA system that accepts questions in one language, L_1, and returns answers in found in another language, L_2. We enumerate below some of these alternatives.

1. One can, for example, use a translation system to translate questions from L_1 into L_2 and an end-to-end question-answering system capable of operating in language L_2. Such a system would contain all stages of a monolingual QA system, from answer type identification and information retrieval through answer selection.
2. Alternatively, one could identify the expected answer type for a question in L_1; use L_1 to generate queries to a CLIR system that searched L_2 using some means other than one-best query translation; then translate L_1 into L_2; and finally perform answer pinpointing in L_2.
3. Or one can translate all L_2 documents into L_1; use an end-to-end question-answering system that operates in language L_1; and then project the answers back into L_2.

Several other alternatives are obviously possible.

From the beginning of our work, we had access to TextMap, a complete end-to-end English QA system that makes extensive use of parsing and syntactic/semantic transformations that have been tuned using large collections of English question-answer pairs [3]. The CL-QA track at CLEF 2003 offered an English document collection with Spanish questions, which tilted our preference towards the first two approaches in the above list. However, before proceeding with an implementation, we thought it would be wise to assess the impact that such a design choice would have on the accuracy of our CL-QA system.

Because the overall accuracy of a QA system is directly affected by its ability to correctly analyze the questions it receives as input, a CL-QA system could be quite sensitive to errors introduced during question translation. In order to assess the potential effect of such errors, we created a development test collection using the QA test collection from the 2002 Text Retrieval Conference (TREC). The first 100 questions from that collection (numbered 1394-1493) were independently translated from English into Spanish by two fluent speakers of Spanish that were also fluent in English. A native speaker of Spanish then reviewed the two sets of translations, selected the better of the two, and corrected a few errors that were observed in that set. The resulting Spanish questions could then be used in conjunction with the answer patterns that were provided with the TREC 2002 QA collection to automatically assess the accuracy of a CL-QA system.[1]

The 100 Spanish questions were given as input to three CL-QA systems.

1. One system (TextMap-SMT) used a Statistical Machine Translation (SMT) system trained on a European Parliament parallel corpus [6] to translate Spanish questions into English; and TextMap to find answers to the translated questions.
2. A second system (TextMap-TMT) used a commercial off-the-shelf Transfer-method Machine Translation (TMT) system (the Systran Web-based system, www.systransoft.com) to translate Spanish questions into English; and TextMap to find answers to the translated questions.
3. A third system (TextMap-English) used a simple lookup table to produce perfect English translations for the Spanish questions in our development collection; and TextMap to find answers to these questions.

Statistical MT systems can be very effective when large quantities of representative translation-equivalent sentences are available for training. Unfortunately, disappointing results were obtained with TextMap-SMT (6/100 correct vs. 35/100 correct with TextMap-English), which led us to abandon this option for CLEF-2003. We believe that at least one reason for these poor results is the difference in genre between the data used to train the statistical MT system (European Parliament Proceedings) and the data used in the context of our evaluation (factoid questions).

TextMap-TMT found correct English answers to Spanish questions about 60% as often as the third system (20/100 correct for TextMap-TMT vs. 35/100 correct for TextMap-English). Our failure analysis process revealed one important improvement that we could make to question translation, however. We observed that Systran translation errors exhibited some clear regularities for certain question types, as might be expected from a rule-based system. For example, Systran exhibited a propensity to produce "whichever" rather than "how many" as a translation for the Spanish word "cuanto." We were able to automatically correct some of these mistakes using a few easily written regular expressions that we developed by inspection on our development collection. Overall,

[1] We provided these translations to the CLEF CL-QA track organizers for use by other teams.

our experiments with the development collection showed that translation quality significantly affects the accuracy of our system.

In order to attribute errors to specific processing stages, we built systems that combined components of TextMap-English and TextMap-TMT. When we used TextMap-TMT only for Answer Pinpointing (described below), 31 of 100 answers were correct (4 below what we get with TextMap-English). When we used TextMap-TMT only for search (both Web Search and Target Collection Search), 28 of 100 answers were correct (7 below what we get with TextMap-English). From this we conclude that while additional work in both cross-language retrieval and answer pinpointing could be useful, improvements in cross-language search may offer somewhat greater potential for improving the overall accuracy of our CL-QA system. We therefore started an effort to incorporate the best available cross-language information retrieval technology into TextMap-TMT; unfortunately, this work could not be not completed before the submission deadline.

3 System Design

The experiments described in the previous section suggested that TextMap-TMT was the best system we could produce under the given time constraints. For the sake of completeness, we briefly review here the main components and data flow in the TextMap-TMT system (see 1 for an illustration of the data flow and [3] for a more detailed presentation of the TextMap system).

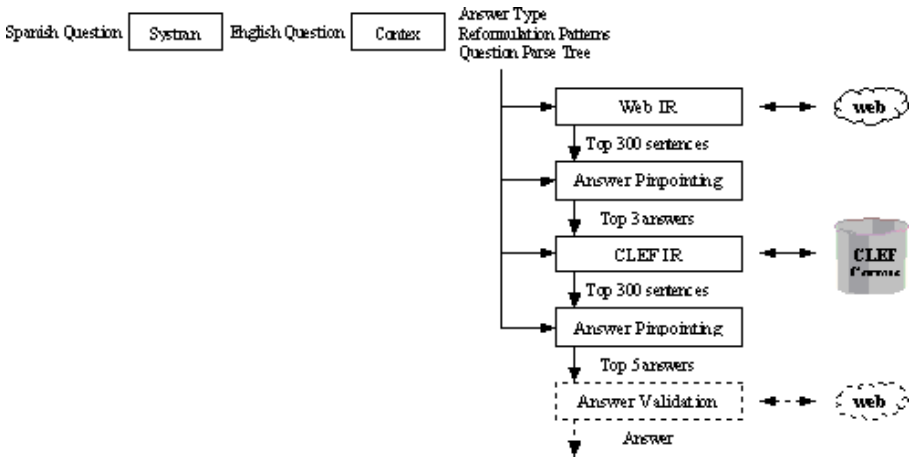

Fig. 1. TextMap-TMT Architecture.

3.1 Using Question Translation for CL-QA

Our TextMap-TMT QA system includes the following major processing stages:

Question Translation. Spanish questions are automatically translated using Systran. A postprocessing module corrects some of the systematic errors that are produced by the translation software, as described above.

Question Analysis. Syntactic analysis is performed using using the CONTEX parser developed at USC-ISI [4], and 143 types of named entities are tagged (e.g., PERSON, ISLAND, SPEED-QUANTITY, BASEBALL-SPORTS-TEAM, and DATE) using an extended version of the BBN IdentiFinder [1].

Answer Type Determination. The results of question analysis are used to automatically classify each question into one of 180 types that describes the expected nature of an appropriate answer (e.g., WHY-FAMOUS, PERSON-NAME, or DISEASE) [5]. Reformulation rules are then used to generate plausible ways of rewriting each question into forms in which an answer might be found (e.g., one rewriting for the question "Where is Devil's Tower?" would be "Devil's Tower, in LOCATION, ...").

Web Search. We use a large set of reformulation patterns and query expansion techniques to produce queries to the Google search engine. For example, the question "¿Cuándo se convirtió Alaska en un estado?" is translated to "When became Alaska a state?" This yields the following Web query:

("Alaska" AND "state" AND "became") OR "Alaska became a state" OR "Alaska became a state on" OR "Alaska became a state in" OR "Alaska became a state about"

Similarly, "Where is Devil's Tower?" yields the following Web query:

("Devil" AND "Tower") OR ("location of Devil's Tower is") OR ("Devil's Tower is located ")

We retrieve the 10 top-ranked documents over the Web, rank each sentence found in that set of documents using a set of locally developed heuristics, and select the top 300 sentences.

Web Answer Pinpointing. The top 300 sentences retrieved by the Web retrieval stage are each then parsed. The reformulation patterns are used here to pinpoint and score answer constituents. Each answer is scored using a wide range of heuristics that measure the degree of overlap between questions and answers; the specificity of the match between the expected answer type and the syntactic/semantic category of the answer; the redundancy of the answer in the answer set; and a number of other factors (see [3] for details).

Target Collection Search. We use the same set of reformulation patterns as Web Search and a similar set of query expansion techniques to produce Inquery queries for the CLEF 2003 CL-QA track English collection (Los Angeles Times articles from 1994). These documents were indexed by Inquery with stopwords retained and the standard stemmer (kstem) enabled. Our use of Inquery for Target Collection Search (rather than MG, which we had previously used) was motivated by the facilities it provides for cross-language search using structured queries, but the cross-language search capability was not ready by the submission deadline. All processing therefore used the same set of translated questions from TMT. The query is expanded by rewarding

the presence of at least one of the top three answers from Web Answer Pinpointing. The question "¿Cuándo se convirtió Alaska en un estado?" yields the query

#passage100(#wsum(10 7 #wsum(49.049 6.730 #1(state) 7.242 #1(became) 35.077 #1(alaska)) 3 #or(#1(January 3 1959) #1(1867) #1(1959))))

We then rank every sentence in the top 1000 documents using the same heuristics as were used in Web search and select the top 300 sentences for use in the Answer Pinpointing stage.

Target Collection Answer Pinpointing. The same Answer Pinpointing module is then applied to the top 300 sentences from the CLEF collection to extract and rank the answers.

Answer Validation. Optionally, we can look back to the Web search results to increase our confidence in some potential answers by giving greater weight to answers that are also present in the top ten sentences found by the Web search stage. As many as five answers can have their confidence raised in this way, and the increase in confidence is proportional to the number of occurrences of that answer in the ten sentences.

Most stages of our original (English) TextMap system could be used unchanged, although we made some relatively minor changes to decouple Web Search from the parsing-based reformulation strategy and to accommodate the presence of Latin-1 characters from above the 7-bit ASCII character set.

4 Results

Figure 2 shows our official results. The center ring depicts the better of the two runs we submitted (which omitted Answer Validation), the top-ranked answer was scored as correct in 56 of 200 cases, 53 of which were judged to be supported by the content of the target collection. As shown on the inner ring, answer validation had the effect of increasing the number of unsupported correct answers (from 3 to 5) and reducing the overall number of correct answers (from 56 to 48). From this we conclude that our present approach to Answer Validation was not helpful in this case.

When used in interactive applications, it can be useful to display multiple candidate answers to the user. Accuracy within the top three answers was therefore chosen as the official measure of effectiveness for the CLEF 2003 CL-QA evaluation. By that measure, shown on the outer ring, our best run found 77 correct answers out of 200, 69 of which were judged to be supported by the content of the target collection.

5 Conclusions and Future Work

The complexity of present QA systems makes cross-language question answering a challenging task; many interacting components must be tuned to work together

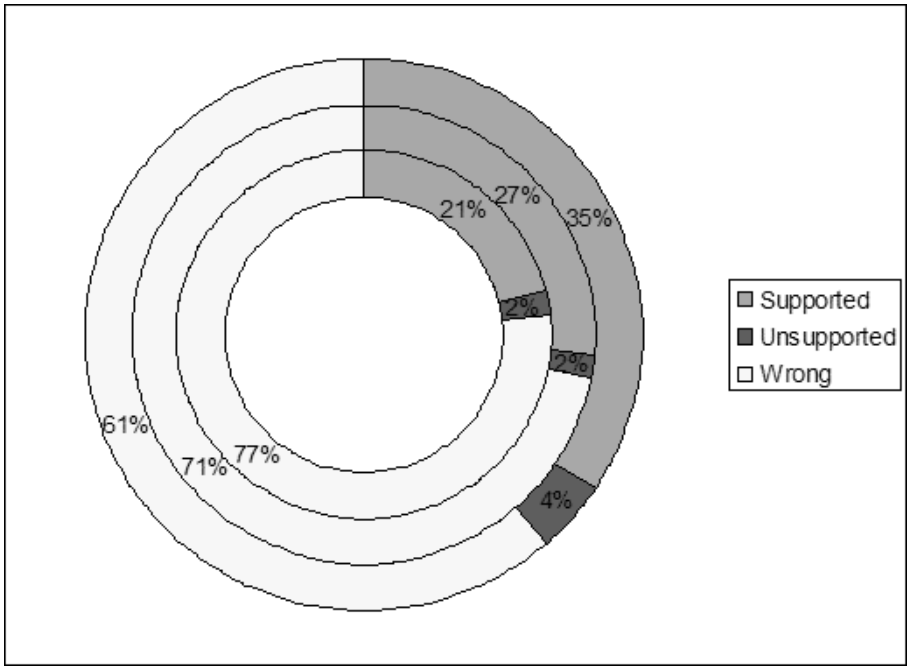

Fig. 2. Official Results. Inner ring: top-1 with Answer Validation; Middle ring: top-1 without Answer Validation; Outer ring: top-3 without Answer Validation.

in new ways. We were therefore pleased to learn that TextMap-TMT found the right answer in better than one out of every four cases, for questions that we had never seen before. Moreover, in the course of our work we identified several promising directions for future work. We observed, for example, that Systran sometimes translates names that should be left untranslated; "Marcos" was translated to "Marks" in our development collection, for example. This would be easily prevented using a Spanish named entity tagger. Another obvious step is to build in more sophisticated approaches to cross-language search than we are presently using.

The most important things we have learned from this experience are not answers, however, they are questions. CL-QA creates a fundamental tension between the design of our best QA systems and the design of our best translation systems. The best QA systems that we are presently able to build rely on relatively deep analysis and a substantial amount of knowledge engineering. These systems are highly tuned to the characteristics we observe in the human use of language. With translation, however, we produce language using a machine. Moreover, the best presently available machine translation systems make extensive use of statistical analysis. If we use these statistics well, there will be no exploitable regularities in their error characteristics (indeed, if there were, then would could have made better use of the statistics!). Put simply, statisti-

cal translation developed using texts in one domain and deep linguistic analysis produced with a system designed to perform well in a different domain don't presently play well together.

Given this fundamental tension, we can see three ways to proceed. One alternative would be to move away from deep linguistic processing and towards greater use of statistical techniques as the basis for our question answering system. This is not an all-or-nothing proposition, of course; all statistics are done on symbols that have linguistic meaning. But as we discover ways of acquiring the large quantities of representative training data that we need, it seems reasonable to expect that the balance will shift in favor of greater reliance on statistical analysis. We have started to explore this direction using noisy-channel models, and we are now achieving results that are competitive with our TextMap system when suitable training data exists [2]. The obvious next step will be to couple this with statistical translation models. Notably, we may not require that our translation models be tuned for fact-based question; it may suffice to tune the model for the typical contents of news stories (essentially, an answer translation strategy).

The obvious alternative would be to perform linguistic analysis in a common framework for both languages; this is essentially what happens inside a transfer method MT system. Versions of our CONTEX parser [4] have been produced for Japanese and Korean as well as English, and parsing for eight languages in a common framework is available from Connexor (www.connexor.com). Ultimately, however, this approach can probably best be explored by those who can leverage the large investments that have already been made in transfer method MT systems. Of course, the best solution will likely ultimately be found somewhere in the middle ground, drawing together the best of the statistical techniques and the deeper linguistic analysis. It may be some time before we can see the shape of these solutions, but our experience at CLEF 2003 has started us on that path.

Acknowledgments

This work has been supported in part by ARDA contract MDA908-02-C-0007 (AQUAINT), DoD contract MDA904-02-C-0406 (LAMP) and DARPA cooperative agreement N660010028910 (TIDES).

References

1. Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1/3), February 1999.
2. Abdessamad Echihabi and Daniel Marcu. A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 7-12 2003.
3. Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu. Natural language based reformulation resource and web exploitation for question answering.

In *Proceedings of the Text Retrieval Conference (TREC–2002)*. November 2002.

4. Ulf Hermjakob. Rapid parser development: A machine learning approach for korean. In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL-2000)*, 2000. http://www.isi.edu/~ulf/papers/kor_naacl00.ps.gz.
5. E.H. Hovy, L. Gerber, U. Hermjakob, C.-Y. Lin, and D. Ravichandran. Toward semantics-based answer pinpointing. In *Proceedings of the DARPA Human Language Technologies Conference (HLT)*, San Diego, CA, 2001.
6. Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. Available at http://www.isi.edu/∼koehn/publications/europarl/, 2003.