Experience with language implementations in ATAMIRI

Presented at the Global Symposium
on Promoting the Multilingual Internet
(International Centre of Geneva, 9-11 May 2006)

Iván Guzmán de Rojas

Abstract

ATAMIRI is a non-commercial system that operates in the Web as a truly **multilingual** machine translator, i.e. one program, one lexical and grammatical data base, supporting various languages capable of operating either as source or target language, with simultaneous translation from any source language to various target languages. The key aspect of this MT technology is its genuinely multilingual property. When a N-th language is implemented, this will immediately be related with the rest of the (N-1) languages in the system. Therefore, implementation costs are only proportional to N. This is an economically significant difference with other systems that try to cover the multilingual demand with multiple programs and dictionaries developed by language pairs therefore with implementation costs proportional to the N(N-1) translation directions in the language set.

This paper describes our operational experience with nine language implementations in ATAMIRI's translator engine: the Latin languages Spanish, French, Portuguese, Italian, Romanian and Catalan and also English, German and Dutch. The resulting 72 language translation directions show various translation quality levels. Both language engineering and economical aspects are discussed. A project plan outline is suggested in order that ATAMIRI technology can be exploited in its full potential.

Multilingual translation demand in the Internet

Although the usefulness of automatic translation on the Internet is now quite well understood and widely available through various commercial on-line services, there remains a large gap between the small number of languages supported and the demand for multilingual translation. For example, the Google Directory lists 75 different languages in the web pages that it indexes suggesting the need in principle to support 5,550 (75x74) different translation directions, but even the most advanced

systems currently support at most 45 language pairs with English and French as preferred source or target languages.

As an illustration consider that nowadays there are in the market machine translation systems for English and German (the two languages with greatest presence in the Web) capable to handle the six Latin languages present in Internet; however, there are lacking translation directions as is shown in the following table:

	English	German	Spanish	French	Italian	Portuguese	Romanian
English	–	YES	YES	YES	YES	YES	NO
German	YES	–	YES	YES	NO	NO	NO
Spanish	YES	YES	–	YES	NO	YES	NO
French	YES	YES	YES	–	NO	NO	NO
Italian	YES	NO	NO	NO	–	NO	NO
Portuguese	YES	NO	YES	NO	NO	–	NO
Romanian	NO	NO	NO	NO	NO	NO	–
Catalan	NO	NO	YES	NO	NO	NO	NO

Rows are source languages, columns are target languages. YES indicates translation direction offered by at least one translation web service or software supplier. Currently there are only 19 YES cases out of the 56 possibilities, all of which are now supported by ATAMIRI in its experimental operation.

The development of an automated translation system using the language pair approach needs enormous financial resources and it demands a considerable R&D effort. For this reason, most of the current developments focus on English.

Translation between Latin languages is not an interesting market for commercial companies; some languages are not even translated into English, like Romanian or Catalan, although Latin languages have particular characteristics in common, like grammatical similarity, which makes its MT development less costly.

Language implementation methodology

- Lexicographic tasks

  Lexicographic enrichment costs are low, since it's possible to centralize the data base management via Internet, while the introduction of new terms is done in a decentralized way, practically from any personal computer connected to Internet. Under the auspices of Unión Latina terminology has been entered from Paris, in collaboration of "Atlas de la Diversidad" the initial basic Catalan lexicon is being introduced from Barcelona, and also some entries in Romanian are being introduced by a volunteer from Bucharest.

  The lexical coding system allows adding lexemes simultaneously in various languages with lexicographic consistency and ensuring integrity of the data base.

  Current number of lexical entries (March 2006)

Spanish	28,106
French	22,574
Italian	15,205
Portuguese	13,660
Romanian	10,109
Catalan	1,564
English	27,387
German	15,836
Dutch	11,466

  The lexicographic data base has also entries in other languages that have not been yet implemented in the translator engine or they are at a very preliminary implementation stage: Aymara (6,393), Russian (9,774), Swedish (2,639) and Hungarian (2,026).

  The lexicographic data base can be viewed at: www.atamiri.cc/arunqera

- Morphological tables

  Word flexion (conjugation of verbs, declination of nouns, adjectives and articles) is handled by ATAMIRI using special morphological tables for each language. Complex flexion rules, for example in Romanian, may take two to three months to develop complete tables. Also contraction rules are handled by tables.

- Syntactic structure entries

  ATAMIRI uses a matrix language representation so that syntagmas of a language are contained in multi-level tables. Our experience with 9 languages shows a requirement of not more than 2,000 syntagmas per language in order to generate well formed sentences of any kind. Syntagmas are manually introduced as they occur during translations. The experiment to use the same syntagmatic tables for Spanish, French, Portuguese and Italian has proven to be positive.

  An explanation of this matrix language representation can be found in:
  New Directions in Machine Translation Conference Proceedings, Budapest 18/19- 8-1988 (John von Neumann Society for Computing Sciences / Dordrecht /Providence: Foris Publishers).

- Translation quality evaluation

  Any method applied to evaluate translation quality has to be designed according to the main purpose of the evaluation. We want to follow the achieved progress during an implementation process verifying how usable the generated translations are. For this simple approach it is enough to consider the intelligibility factor obtained in a translated text as an average of the corresponding factors assigned to each sentence of the text after

reading it. This method is explained with examples at the Aynisiwi forum: www.atamiri.cc/aynisiwi in the lexicographic group section.

Please see:

Aynisiwi -> Lexicographical workgroup -> Catalan preliminary implementation -> L'AMETLLER -> Evaluation of the English translation

Aynisiwi -> Lexicographical workgroup -> Evaluation Results

Conclusions

Experience with 9 languages implementations in ATAMIRI's translator engine shows the following facts:

- A multilingual translator engine is feasible and capable to operate in Internet at high speed. In spite of the difficulties to develop a common language representation in a multilingual environment, until now ATAMIRI has shown that it is possible to apply a kind of universal grammar valid for languages with different linguistic characteristics.

- An initial implementation stage for a language requires 4 months of work with 2 fully dedicated persons. At the end of this first stage we can expect usable translations but in a very restricted semantic field with a limited collection of texts. The minimum required lexicon is 3 to 4 thousand entries. The main purpose of this initial work is to test ATAMIRI's capability to handle the grammar of the language being implemented and to detect the need of further program adjustments. For example, we have now Portuguese, Italian, Romanian, Catalan, German and Dutch at this stage with positive preliminary results.

- In a second implementation stage we pursue to evaluate quality translation of texts within a chosen thematic. With help of the

lexicographic analyzer we introduce the required lexicon for those texts. The minimum required lexicon is 20,000 entries. During this stage we found that translations of postings in the Aynisiwi virtual forum became quite intelligible and usable. The main purpose here was to enrich the syntagmatic data base of the language being implemented and also to introduce the necessary program adjustments to solve unexpected homograph cases. At this stage currently we have English, Spanish and French.

- The third implementation stage, where we pursue to improve translation quality to higher levels, we have not been able to start yet due to lack of the required financial resources. At this stage we should reach at least 40,000 lexicon entries per language in order to run extensive translation tests with a wide collection of texts reading web pages in various fields of contents. As a result we expect to achieve average intelligibility factors greater than 8.5 in all implemented translation directions.

- One important result obtained from the operation run so far is our differentiated knowledge of all types of generated anomalies which affect translation quality. We have found that almost all of these cases can be solved within the frame of the original design of ATAMIRI's translation engine; therefore we are confident that a third implementation stage could lead to obtain the expected results.

- Another result of this experience is the realistic understanding gained of the cost factors involved in language implementations using ATAMIRI. Roughly speaking, all three implementation stages, without considering initial R&D costs of the translator engine, have a global cost of 200,000 US$ per implemented language. The three implementation stages share this amount in 20%, 40% and 40% respectively.

Project Plan Outline

World wide cooperation is required to mobilize the competencies needed to address the multilingualism issue; otherwise the Internet will remain as a collection of isolated linguistic islands, missing the opportunity of direct communication within cultural diversity. As the creator of ATAMIRI, I urge leaders of institutions and corporations that promote Language Engineering projects and government authorities concerned with the problematic of Human Language Technologies, to support a thorough ATAMIRI assessment operation to test its multilingual technology and verify its translation quality improvement capacity.

ATAMIRI's R&D is still advancing, even though too slowly. It is waiting for its technological potential to be exploited on large scale. To achieve this, capital investment, under equitable conditions that recognizes the value of this unique technology, is required. We propose to:

- ➔ Create an ATAMIRI Language Engineering Network (ALEN) dedicated to further develop and exploit ATAMIRI's technology in its full potential.

- ➔ Transfer ATAMIRI technology to ALEN with the system creator as on of its stakeholders. The stakeholders have the option to adopt an open source model for the software.

- ➔ Obtain funding commitments from various stakeholders to ensure an initial and continuous five year operation offering free Internet translation services for messenger, email and virtual forums, completing all five implementation stages for the languages used so far and for few additional ones of high priority for the stakeholders.

- ➔ Offer paid Internet translation services of large documents in those languages for which the system is already capable to generate high quality translations.