interNOSTRUM:

A Spanish-Catalan machine translation system

Written by:

Mikel L. Forcada
mlf@dlsi.ua.es

This article describes a Spanish-Catalan machine translation system, interNOSTRUM. The main reason for the demand for translations from Spanish (the official language of Spain) into Catalan is the impulse toward `linguistic normalization' in the Catalan-speaking regions (10 million inhabitants, about 6 million speakers) where Catalan was receding and where the language is now co-official. The interNOSTRUM system is currently under development and a prototype has just started to serve the Universitat d'Alacant, a medium-sized university, and the Caja de Ahorros del Mediterráneo, one of the largest Savings Banks in Spain; these are the two institutions who started and currently fund this three-year project (1999-2001), which has a staff of two linguists (Amaia Iturraspe-Bellver and Anna Esteve) and three computer engineers (Raül Canals, Alícia Garrido and Sergio Ortiz). Even though translation accuracy and vocabulary coverage can still be much improved, the speed of the system --thousands of words per second or millions of words per day on a 1999 model desktop machine acting as an Internet server-- has prompted its use as a system to obtain instantaneous rough translations that are relatively easy to correct into publishable documents. These speeds are achieved through the use of finite-state technology in most of its modules.

The current prototype and the future versions

As has been said, the current version of interNOSTRUM is not a finished product, but rather a prototype that reflects its current state of development, which can nevertheless be used to obtain instantaneous rough translations ready for post-editing. Indeed, two of the basic objectives of our project have been, first, to generate an operational version of interNOSTRUM as soon as possible (launched November 1999) and, second, to make the latest stable version available as soon as it is ready. These are the main reasons for its current configuration as a single Internet server (soon with a mirror for internal use at the Caja de Ahorros del Mediterráneo).

Currently, interNOSTRUM only translates unformatted ANSI or ASCII texts from Castillian Spanish to the central or Barcelona variety of Catalan (the version generating a Valencia variety and a Balearic Island variety will be ready by the end of the project), but both an RTF (Microsoft's Rich Text Format) and an HTML (HyperText Markup Language) version are about to be launched. We expect to release the inverse (Catalan-Spanish) translator by October 2000.

Computational details

The machine translation engine currently runs on the Linux operating system and may be accessed through an Internet server (http://www.tor-simany.ua.es). It consists of 6 independent subprograms that run in parallel and communicate through text channels. The use of human-readable text channels (pipeline) alows for an easy diagnosis of many problems and is a very efficient alternative in Linux (Unix) implementations. Each subprogram is automatically generated from the corresponding linguistic data – a feature that makes interNostrum easily extensible fo other languages - using compilers written with the aid of yacc and lex, which are standard in Unix environments. (The Linux versions are called bison and flex). The current speed of the system is in the order of thousands of words per second on a standard 1999 desktop PC (based on a Pentium at 400 MHz).

Linguistic design

interNOSTRUM is a classical indirect machine translation system using an advanced morphological transfer strategy very similar to the one used in commercial PC-based machine translation systems such as Transparent Technologies' Transcend RT, early versions of Globalink's Power Translator, and Softissimo's Reverso. interNOSTRUM's operation has the following stages:

1. ANALYSIS:
   - Morphological analysis
   - Part-of-speech tagging
2. TRANSFER:
   - Bilingual dictionary lookup
   - Word pattern processing (agreement, reordering lexical changes)
3. GENERATION:
   - Morphological generation
   - Postgeneration (rules for Catalan apostrophes and hyphens)

Subprograms based on finite-state technology

Four of the subprograms in interNOSTRUM, namely, the morphological analysis, the bilingual

dictionary lookup and the morphological generation and the postgeneration program are based on finite-state transducers, a technology that allows for processing speeds in the order of thousands of words a second, speeds that are practically independent of the size of the dictionaries. Finite-state transducers read their input symbol by symbol; each time a symbol is read, they move to a new state, and write, also symbol by symbol, one or more output symbols.

Morphological analysis: The morphological analysis program, which is automatically generated from a morphological dictionary for the source language (SL). The morphological dictionary contains the lemmas (canonical or base forms for inflected words), the inflection paradigms, and their mutual relationships. The subprogram reads the text or surface forms and writes, for each surface form, one or more lexical forms consisting of a lemma, a part of speech, and inflection information.

Bilingual dictionary lookup: The bilingual dictionary lookup subprogram is called by the pattern processing subprograms (see below); it is automatically generated from a file that contains the bilingual correspondences. The program reads a source-language lexical form and writes the corresponding target-language lexical form.

Morphological generation: Morphological generation is basically the reverse of morphological analysis, but applied to the target language. The morphological generation program is generated from a morphological dictionary for the target language.

Postgeneration: The surface forms involved in apostrophizing and hyphenation (such as clitic pronouns, articles, some prepositions, etc.) activate this subprogram which is otherwise asleep. The postgenerator is generated from a simple file containing the corresponding rules for the target language.

The division of a text in words has some nontrivial aspects; two of them will be mentioned: the problem of some multiword units and that of Spanish enclitic pronouns.

Multiword units: There are a number of word groups that cannot be translated word for word and may be treated as fixed-length multiword units; they are currently being incorporated into the bilingual dictionary:

```
• Spanish con cargo a Catalan a càrrec de ("at the expense of'')
• Spanish por adelantado Catalan per endavant ("in advance'')
• Spanish echar de menos Catalan trobar a faltar ("to miss (someone)'')
```

In the last example, the multiword unit has a variable element that may be inflected (in boldface); multiword units with inflection have just started to be incorporated into interNOSTRUM's dictionaries.

Enclitic pronouns: The morphological analysis program is also able to solve combinations of certain verb forms and enclitic pronouns, which are written in Spanish as a single word; these combinations occur with orthographical transformations such as accent marks or loss of consonants:

```
• Sp. dámelo = da + me + lo   Cat. dóna + me + lo = dóna-me'l ("give it to me!'')
• Sp. presentémonos = presentemos + nos   Cat. presentem + nos = presentem-nos ("let us introduce ourselves'')
```

The lexical disambiguation program

Lexical ambiguities fall into two main groups: homography (when a surface form has more than one lexical form or analysis) and polysemy (when the surface form has a single lexical form but the lemma may have more than one interpretation).

The lexical disambiguation subprogram --very similar to a part-of-speech tagger-- uses a language model based on trigrams (sequences of three lexical categories) to solve a fraction of the homographs ocurring in Spanish texts. The model's parameters reflect the frequencies observed for each trigram in a reference text corpus, and assigns a probability to each possible disambiguation of a sentence containing a lexical categorial ambiguity and the most likely disambiguation is chosen. The current performance of this subprogram is suboptimal because of the insufficient statistical significance of the text corpora used in the first prototype. We are currently fine-tuning the tagset used and building a larger corpus for retraining to improve the performance of this subprogram.

The few errors occurring in certain difficult but frequent homographs, such as una (article/verb unir, observed frequency 0.0077), para (verb parar) /preposition, observed freq. 0.0077), and como (conjunction/verb comer), observed freq. 0.0043) constitute one of the main contributors to the current error rate in interNOSTRUM. Fortunately, other frequent homographs are not so difficult to disambiguate.

Lexical ambiguities occurring inside the same lexical category are currently tackled by means of ad hoc strategies. A fine-tuning of the tagset will solve some homographs. Rules regarding polysemic words will be included in a controlled Spanish biased toward banking and administration applications, which will include syntactical restrictions in addition to the lexical ones. A style assistant will help authors follow the rules of controlled language when preparing Spanish documents.

The pattern processing subprogram

In spite of the great similarity between Spanish and Catalan, there are still a number of important grammatical divergences:

• modal constructions: Spanish tienen que firmar Catalan han de signar ("they have to sign");

• gender and number divergences: Spanish la deuda contraída (fem.) Catalan el deute contret

(masc.; English "the assumed debt");

• dropping of prepositions before que: Spanish la intención de que el cliente esté satisfecho Catalan la intenció que el client estigui satisfet (Engl. "the intention that the customer be satisfied");

• relative constructions using cuyo ("whose"), absent in Catalan: Spanish la cuenta cuyo titular es el asegurado Catalan el compte el titular del qual és l'assegurat (English "the account whose owner is the insured person").

These divergences have to be treated using suitable grammatical rules.

interNOSTRUM uses a solution which may also be found in commercial MT systems. It is based on the detection and treatment of predefined sequences of lexical categories (patterns) which may be seen as rudimentary phrase-structure constructs: for example, *art.-noun* or *art.-noun-adj.* are two possible valid noun phrases. Those sequences known to the program constitute its pattern catalog. The subprogram follows a pattern-action scheme as follows:

• The text (morphologically analysed and disambiguated) is read left to right, one lexical category at a time.

• The subprogram searches, starting at the current position in the sentence, for the longest sequence that matches a pattern in its pattern catalog (for example, if the text starting in the current position is "un senyal inequívoc..." ("an unmistakable signal"), it will choose *art.-noun-adj.* instead of *art.-noun*).

• The subprogram operates on this pattern (to propagate gender and number agreement, to reorder it, to make lexical changes) following the following the rules associated to the pattern.

• Once the operation has finished, the pattern processing subprogram continues immediately after the pattern just processed (it does not re-visit any of the words on which it has already operated).

When no pattern is detected in the current position, the program translates one word literally and restarts at the following word. "Long-range" phenomena such as subject-verb agreement are harder to treat; they require the propagation of information from one pattern to the following ones. We are currently working on this aspect.

The pattern processing subprogram is automatically generated from a file containing rules that specify the patterns and the associated actions. This is the slowest subprogram; whereas the rest of the subprograms work at tens of thousands of words per second, the speed of the pattern processing program is in the order of one thousand words per second. Its current pattern catalog only contains a few of them.

Projected support tools for interNOSTRUM

We are currently working on the following support tools:

• A style assistant which will help the author of a Spanish text avoid many difficult ambiguities using the syntactical, lexical and style rules specified in a controlled Spanish.

• A pre-editing assistant, that will allow for a manual disambiguation of problematic words and structures, simply clicking on them to get a menu of options. This will be very helpful in those cases in which the statistical strategy used by the program is incapable of making the right choice.

• A post-editing assistant, in which the author will be able to click on a target-language word when he or she suspects that it is an incorrect translation and will allow him or her to substitute it by an alternative, taking into account the original text. ■