An open source shallow-transfer machine translation engine for the romance languages of Spain

A.M. Corbí-Bellot[1], M.L. Forcada[1], S. Ortiz-Rojas[1], J.A. Pérez-Ortiz[1],
G. Ramírez-Sánchez[1], F. Sánchez-Martínez[1], I. Alegria[2], A. Mayor[2], K. Sarasola[2]

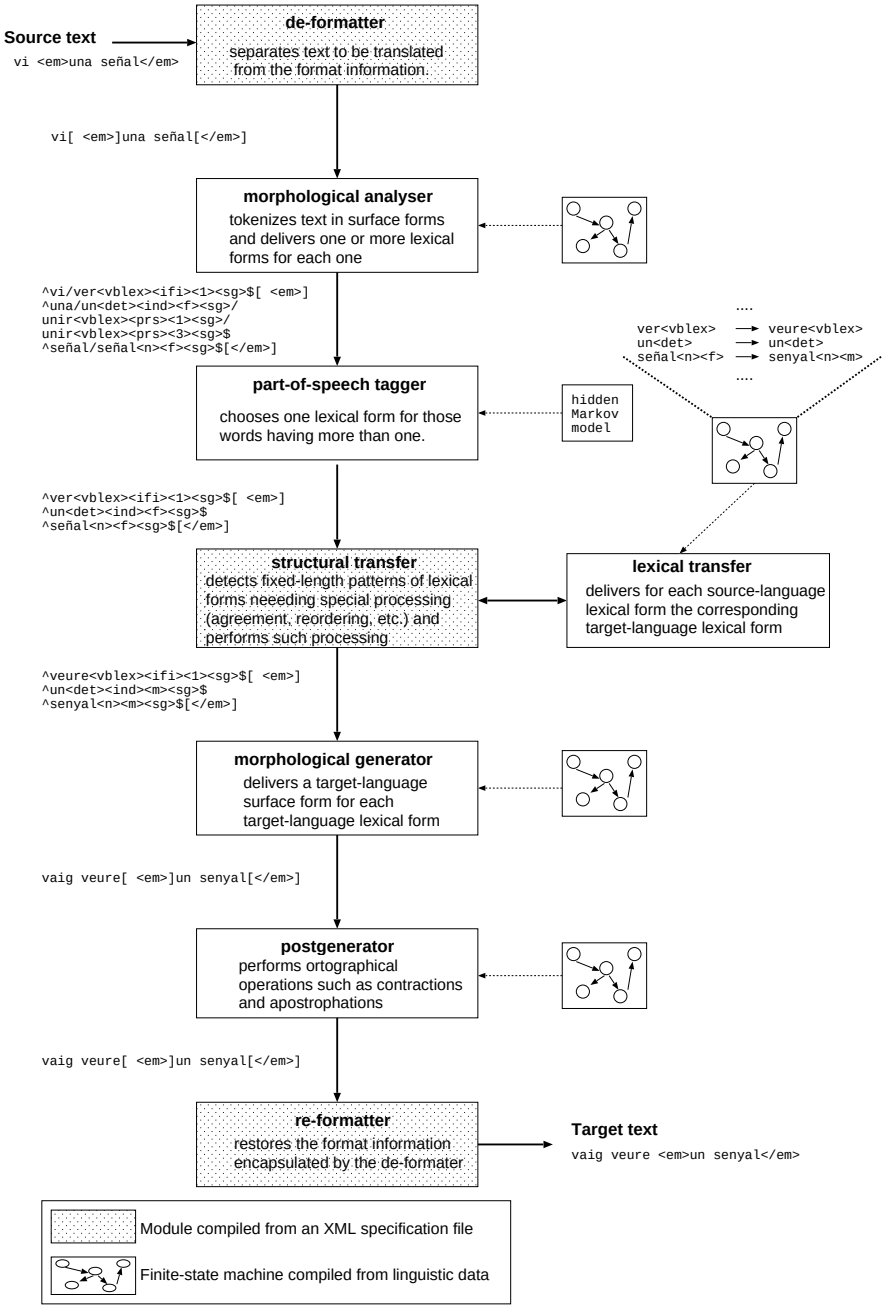
[1]Transducens Group, Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03071 Alacant, Spain
[2]IXA Taldea, Informatika Fakultatea, Euskal Herriko Unibertsitatea, E-20071 Donostia, Spain

Why?

- The multilingual nature of Spain is recognised in laws and regulations corresponding to the various levels of government.
- Demand from many citizens make private companies to be interested in generating information in languages different from Spanish.
- The various levels of government must respect the linguistic rights recognised to their citizens and promote the use of such languages.
- Existing MT programs for Spanish–Catalan and Spanish–Galician are mostly commercial and use proprietary or closed technologies.
- Closed and proprietary technologies are very hard to adapt to new usages.

Goals

- To release a shallow-transfer machine translation engine under an open-source license, together with pilot linguistic data.
- To introduce a unified open-source MT architecture to solve problems caused by proprietary and closed technologies.
- To propose XML-based standards to code the linguistic data, making it easier to adapt the system to new language pairs.



The 8-module MT system architecture

Format for linguistic data

(See `http://torsimany.ua.es/eamt2005/`)

Dictionaries (lexical processing)

- Morphological and bilingual dictionaries are an XML evolution of the format already used in interNOSTRUM [**?**] and Traductor Universia [**?**].
- Morphological dictionaries establish the correspondence between surface forms and lexical forms.
- Bilingual dictionaries establish correspondences between source-language and target-language lexical forms.
- Post-generation dictionaries establish correspondences between input and a output strings corresponding to orthographical transformations.

Tagger definition

- Source-language lexical forms are defined in terms of fine part-of-speech tags which are necessary in some modules.
- For the PoS tagging purpose is convenient to define a coarser tagset grouping fine tags into coarser ones.
- The tagset definition file is an XML file defining coarse tags and some additional information useful for the PoS tagging such as forbidden tag sequences.

Structural transfer

- The transformation to be performed are defined in an XML file containing patterns and actions for each detected pattern.

De-formatter and re-formatter

- For each supported format, the de-formatter and re-formatter are specified through the same XML file.
- The XML definition file defines those rules needed to encapsulate the format information.

Conclusions

- An open-source shallow-transfer MT engine for the Romance languages of Spain has been described.
- The presented MT engine can be adapted to translating between other Romance languages not considered in the project, or for other non-Romance, but closely related languages, such as Swedish–Danish or Czech–Slovak.
- Some components of this architecture can also be used for other natural language processing tasks.
- The code, together with pilot Spanish–Catalan and Spanish–Galician linguistic data will be released at the beginning of 2006.

References

[1] Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M.I. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón, M.L. Forcada (2001). "The Spanish-Catalan machine translation system interNOSTRUM", in B. Maegaard, ed., *Proceedings of MT Summit VIII: Machine Translation in the Information Age*, 73–76.

[2] Garrido-Alenda, A., Patrícia Gilabert Zarco, Juan Antonio Pérez Ortiz, Antonio Pertusa-Ibáñez, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Míriam A. Scalco, Mikel L. Forcada (2004). "Shallow parsing for Portuguese-Spanish Machine Translation", in Branco, A. and Mendes, A., Ribeiro, R., *Language technology for Portuguese: shallow processing tools and resources*, 135-144.