Shallow Parsing for Portuguese–Spanish Machine Translation

Alicia Garrido-Alenda, Patrícia Gilabert-Zarco, Juan Antonio Pérez-Ortiz, Antonio Pertusa-Ibáñez, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Miriam A. Scalco and Mikel L. Forcada

Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain.
mlf@ua.es

Abstract

To produce fast, reasonably intelligible and easily corrected translations between related languages, it suffices to use a machine translation strategy which uses shallow parsing techniques to refine what would usually be called *word-for-word* machine translation. This paper describes the application of shallow parsing techniques (morphological analysis, lexical disambiguation, and flat, local parsing) in a Portuguese–Spanish, Spanish–Portuguese machine translation system which is currently being developed by our group and is publicly and freely available at `http://copacabana.dlsi.ua.es`.

1. Introduction

We describe the successful application of shallow parsing techniques in a Portuguese–Spanish, Spanish–Portuguese machine translation (MT) system which is currently being developed by our group and is publicly and freely available at `http://copacabana.dlsi.ua.es`.

The paper is organized as follows: section 2 describes the role of shallow parsing in real-world related-language machine translation. The Portuguese–Spanish MT engine is described in section 3. Lexical disambiguation and structured is discussed with a bit more detail in sections 4 and 5. Section 6 ends the paper with a few concluding remarks.

2. Real Machine Translation and Shallow Parsing

General-purpose MT systems are expected to satisfy the requirements of the two main application modes: *assimilation* or understanding of documents written in another language (fast, intelligible translations) and *dissemination* of documents translated into another language (easily correctable translations).

Real (i.e., working) MT may be seen both as the result of approximations (some of them inevitable) over an ideal, theoretically motivated model based on the *principle of semantic compositionality* and as the result of a set of necessary refinements over a very rudimentary *word-for-word* substitutional system.

On the one hand, real MT may be seen as a set of successive approximations over "ideal MT":

1. Most MT system adopt the approximation that *translating texts is translating sentences*, which, for example, excludes the treatment of some aspects of discourse structure.
2. The *principle of semantic compositionality* (PSC, Radford et al. 1999, p. 359) states that the interpretation (meaning) of a sentence is compositionally built from the interpretation of its words, following the groupings dictated by its parse tree, and also conversely, sentences may be compositionally built from interpretations (Tellier, 2000). Translating a source language (SL) sentence would then mean (a) *fully parsing* it, (b) assigning interpretations to its words, (c) compositionally building an interpretation, (d) analysing this interpretation to obtain target language (TL) words and a TL parse tree from it, and (e) generating a TL sentence from them. This is basically the *modus operandi* of interlingua systems and constitutes the *compositional translation* approximation. Note that this account assumes that *lexical* ambiguity (words having more than one interpretation) and *structural* ambiguity (sentences having more than one parse tree) have been also ideally solved.
3. As is the case with professional translators, MT systems do not always need to completely "understand" (build explicit interpretations of) SL sentences. *Transfer* systems take a shortcut and go from SL parse tree and words directly into TL parse tree and words: they do so by applying parse tree transformations (*structural transfer*) and word substitutions (*lexical transfer*), without building an explicit representation of the interpretation. This constitutes an additional approximation, the *transfer approximation*.
4. When languages are syntactically similar (e.g, when related), full parsing is not performed; lexical transfer is complete, but structural transfer is partial and local and occurs only where required. This could be called the *partial parsing* approximation. *Transformer* systems (Arnold et al., 1994, 4.2), many of them commercial and available on the internet[1], are an example of this approximation.

On the other hand, real MT may be seen as a refinement over what would usually be called *word-for-word* MT (which processes input one word at a time and substitutes it by a constant equivalent independently

[1] For example, SDL Transcend is available through `http://www.freetranslation.com` and Reverso is available as `http://www.reverso.net`.

of context). Taking the previous experience of our research group with the interNOSTRUM (`http://www.interNOSTRUM.com`) Spanish–Catalan MT system (Canals-Marote et al., 2001), used by hundreds of people on a daily basis, we can state that, to produce fast, reasonably intelligible and easily corrected translations between related languages —such as Portuguese (`pt`) and Spanish (`es`)—, it suffices to augment *word-for-word* MT with a robust *lexical* processing (to treat multiword expressions and to adequately choose equivalents for lexically ambiguous words), and a local *structural* processing based on simple and well-formulated rules for some simple structural transformations (reordering, agreement).

These requirements are very well met by *shallow parsing* techniques, which are usually applied sequentially:

1. tokenization and morphological analysis, to be able to build bilingual dictionaries as correspondences between SL and TL lemmas, to be able to identify multiword expressions and to determine the syntactic role of each word in the sentence;
2. categorial disambiguation (to choose among multiple analyses in the case of homographs), and
3. partial, flat parsing of those structures needing treatments that may be applied locally.

The next section illustrates how these operations are integrated into the complete dataflow of a `pt–es` machine translation system.

3. The `pt–es` Machine Translation Engine

As said above, we are currently developing a bidirectional MT system between `pt` and `es` (prototype available at `http://copacabana.dlsi.ua.es`) with emphasis in Brazilian `pt`, based on an existing Spanish–Catalan MT system. The current text coverage surpasses 95%, errors rate below 10%, and speed surpasses 5000 words per second on an desktop PC equipped with an AMD 2100 processor. The system, which already receives thousands of visits a day, (a) translates ASCII, RTF and HTML documents and e-mail messages, (b) translates Internet documents (webpages) during browsing, with link following, and (c) implements a bilingual chat room.

The translation engine is a classical *partial transfer* or *transformer* system consisting of an 8-module *assembly line*; to ease diagnosis and testing, these modules communicate between them using text streams. Five modules are automatically generated from linguistic data files using suitable compilers. The modules (organized as in figure 1) are:

- The *unformatter* separates the text to be translated from the format information. Format information is encapsulated so that the rest of the modules treat it as blanks between words.
- The *morphological analyser* tokenizes the text in surface forms (SF) (lexical units as they appear in texts) and delivers, for each SF, one or more lexical forms (LF) consisting of *lemma*, *lexical category* and morphological inflection information. Tokenization is not straightforward due to the existence, on the one hand, of contractions (e.g., *daquele* = *de* + *aquele* [“of that”]), and, on the other hand, of multiword lexical units (*no entanto* [“in spite of”]), which may inflected (*dava na vista* [“called someone’s attention”]). This module is compiled from a SL morphological dictionary (MD) (Garrido et al., 1999; Garrido-Alenda et al., 2002). For example, the `pt` input “`as viagens coletivas`” would give a sequence of four LF’s, with the first one being ambiguous: (*o*, article, feminine plural) and (*o*, clitic pronoun, feminine plural), (*viagem*, noun, feminine plural), and (*coletivo*, adjective, feminine plural).
- The *categorial disambiguator* (part-of-speech tagger) chooses, using a hidden Markov model (HMM) trained on representative SL texts, and according to its context, one of the LFs corresponding to an ambiguous SF. Ambiguous SFs are a very frequent source of errors when incorrectly solved. In the example above, the system would choose (*o*, article, feminine plural), (*viagem*, noun, feminine plural), and (*coletivo*, adjective, feminine plural).The *lexical transfer* module is called by the structural transfer module (see below); it reads each SL LF and delivers the corresponding TL LF. This module is compiled from a bilingual dictionary. In the example, the SL LFs are translated to (*el*, article, feminine plural), (*viaje*, noun, **masculine** plural) — note the gender change — , and (*colectivo*, adjective, feminine plural).
- The *structural transfer* module uses finite-state pattern matching to detect (in the usual left-to-right, longest-match way) patterns of LFs (phrases) needing special processing due to grammatical divergences between the two languages (gender and number changes, reorderings, lexical changes, etc.) and performs the corresponding operations. This module is compiled from a transfer rule file (Garrido-Alenda and Forcada, 2001), and generates a `lex` (Lesk, 1975) scanner as an intermediate step during compilation. In the running example, the noun phrase pattern *article–noun–adjective* is detected; this pattern dictates that the article and the adjective should agree with the translation of the noun, producing: (*el*, article, masculine plural), (*viaje*, noun, masculine plural), and (*colectivo*, adjective, masculine plural).
- The *morphological generator* delivers a TL SF for each TL LF, by suitably inflecting it. This module is compiled from a TL MD. In our example, the result would be the text “`los viajes colectivos`”.
- The *postgenerator* performs orthographical operations such as contractions (*de* + *el* = *del*, etc.) and is compiled from a rule file.
- The *reformatter* restores the original format information into the translated text.

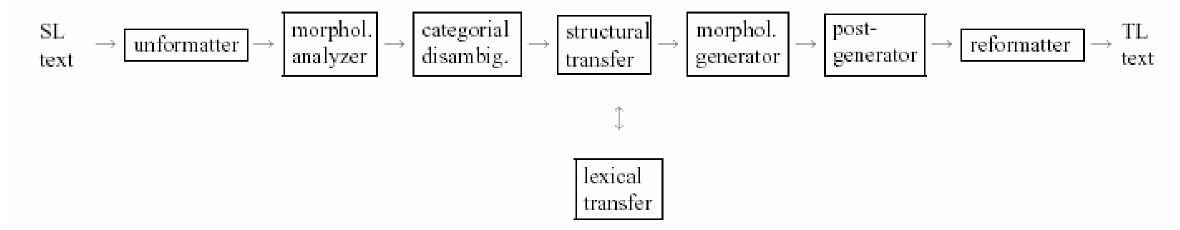

Figure 1: The eight modules of the `pt–es` machine translation system (see section 3).

The morphological analyser, lexical transfer module, morphological generator, and postgenerator are all based on finite-state transducers (Garrido et al., 1999; Garrido-Alenda et al., 2002).

4. Lexical Disambiguation

Building a lexical disambiguator (part-of-speech tagger) based on HMMs (Cutting et al., 1992) for the SL in a MT system implies: (a) designing or adopting a reduced tagset (set of parts of speech) which groups the finer tags delivered by the morphological analyser into a small set of coarser tags adequate to the translation task; (b) building a representative SL training corpus and manually tagging a portion of it for training (in the case of supervised training) and evaluation; (c) actually training the hidden Markov model on the corpus to obtain the probabilities.

After having used for `pt` the disambiguator (tagset and probabilities) developed for Spanish–Catalan (a choice which was adequate for initial prototypes), we have just deployed a new `pt` disambiguator designed as mentioned above.

The tagset used by the `pt` lexical disambiguator consists of 122 coarse tags (83 single-word and 39 multi-word tags for contractions, etc.) grouping the 2230 fine tags (365 single-word and 1845 multi-word tags) generated by the morphological analyser. The number of different lexical probabilities in the HMM is drastically reduced by grouping words in ambiguity classes (Cutting et al., 1992) receiving the same set of part-of-speech tags: 303 ambiguity classes result. In addition, a few words such as *um* (indefinite article or pronoun) or *ter* (to have, auxiliary verb or lexical verb) are assigned special hidden states. The current disambiguator has been trained as follows: initial parameters are obtained in a supervised manner from a 20,000-word hand-tagged text and the resulting tagger is retrained (using Baum-Welch reestimation as in Cutting et al., 1992) in an unsupervised manner over a 7,800,000-word text. Using an independent 6,600-word hand-tagged text, the observed coarse-tag error rate is 4.89%, with about half of the errors (2.14%) coming from words unknown to the morphological analyser[1].

5. Shallow Parsing for Structural Transfer

Many of the structural transfer rules in the Spanish–Catalan system are used without change for `pt–es`: mainly, all rules ensuring gender and number agreement for about twenty very frequent noun phrases (determinant–noun, determinant–noun–adjective, determinant–adjective–noun, numeral–noun etc.), as in *um sinal vermelho* (`pt`, masc.) [“a red signal”]) ! *una se˜nal roja* (`es`, fem.). In addition, we have rules to treat very frequent `pt–es` transfer problems, such as these:

- Rules to choose verb tenses; for example, `pt` uses the subjunctive future (*futuro do conjuntivo*) both for temporal and hypothetical conditional expressions (*quando vieres* [“when you come”], *se vieres* [“if you came”]) whereas `es` uses the present subjunctive in temporal expressions (*cuando vengas*) but imperfect subjunctive for conditionals (*si vinieras*).
- Rules to rearrange clitic pronouns (when enclitic in `pt` when proclitic in `es` or vice versa): *enviou-me* (`pt`) ! *me envi´o* (`es`) [“he/she/it sent me”]; *para te dizer* (`pt`)!*para decirte* (`es`) [“to tell you”], etc.
- Rules to add the preposition *a* in some modal constructions (*vai comprar* (`pt`) ! *va a comprar* (`es`) [“is going to buy”]).
- Rules for comparatives, both to deal with word order (*mais dois carros* (`pt`) ! *dos coches m´as* (`es`) [“two more cars”]) and to translate *do que* (`pt`) [“than”] as *que* (`es`).
- Lexical rules, for example, to decide the correct translation of the adverb *muito* (`pt`) ! *muy*/*mucho* (`es`) [“very”, “much”] or that of the adjective *primeiro* (`pt`)! *primer*/*primero* (`es`) [“first”].

The rules are written in a high-level language (Garrido-Alenda and Forcada, 2001) in the usual *pattern–action* format of `lex`, where the pattern describes the LFs constituting the chunk which is processed and the action performs the actual transformation of the pattern, with lexical transfer always being implicitly called. The resulting module works left to right, processing always the input prefix of the remaining text which matches the longest pattern, and continuing immediately after the pattern. When input does not match any of the patterns, a LF is translated in isolation and processing continues after it. Left-to-right “state” information may be used to

[1] In the current version, 4.40% of the words were unknown to the morphological analyser

communicate the information computed during processing of a chunk to other chunks following it.

6. Concluding Remarks

The speed (5600 words/s on a regular desktop PC) and accuracy (around 90%) mentioned above confirm that the shallow-parsing-based strategy previously used by our group to build a Spanish–Catalan MT system is also adequate for `pt–es` MT.

Acknowledgements: Work funded by Portal Universia, S.A. and partially supported by the Spanish Comisión Ministerial de Ciencia y Tecnología through grant TIC2000-1599-CO2-02.

7. References

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., and Sadler, L. (1994). *Machine translation: An introductory guide*. NCC Blackwell, Oxford. available at `http://clwww.essex.ac.uk/~doug/MTbook/`.

Canals-Marote, R., Esteve-Guillen, A., Garrido-Alenda, A., Guardiola-Savall, M., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Perez-Antón, P., and Forcada, M. (2001). The Spanish-Catalan machine translation system interNOSTRUM. In *Proceedings of MT Summit VIII: Machine Translation in the Information Age*, pages 73–76. Santiago de Compostela, Spain, 18–22 July 2001.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In Third Conference on Applied Natural Language Processing. Association for Computational Linguistics. Proceedings of the Conference., pages 133–140, Trento, Italy.

Garrido, A., Iturraspe, A., Montserrat, S., Pastor, H., and Forcada, M. L. (1999). A compiler for morphological analysers and generators based on finite-state transducers. *Procesamiento del Lenguaje Natural*, (25):93–98.

Garrido-Alenda, A. and Forcada, M. L. (2001). Morphtrans: un lenguaje y un compilador para especificar y generar módulos de transferencia morfológica para sistemas de traducción automática. *Procesamiento del Lenguaje Natural*, 27:157–164.

Garrido-Alenda, A., Forcada, M. L., and Carrasco, R. C. (2002). Incremental construction and maintenance of morphological analysers based on augmented letter transducers. In *Proceedings of TMI 2002 (Theoretical and Methodological Issues in Machine Translation, Keihanna/Kyoto, Japan, March 2002)*, pages 53–62.

Lesk, M. (1975). Lex—a lexical analyzer generator. Technical Report Technical Report 39, AT&T Bell Laboratories, Murray Hill, N.J.

Radford, A., Atkinson, M., Britain, D., Clahsen, H., and Spencer, A. (1999). *Linguistics: An introduction*. Cambridge Univ. Press, Cambridge.

Tellier, I. (2000). Semantic-driven emergence of syntax: the principle of compositionality upside-down. In *Proc. 3rd Conference on the The Evolution of Language*, pages 220–224, Paris.