Learning machine translation strategies using commercial systems: discovering word-reordering rules[*]

Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain.
E-mail: `mlf@dlsi.ua.es`
URL: `http://www.dlsi.ua.es/~mlf`

Abstract

Commercial machine translation (MT) systems seldom publicize their MT strategies; however, unlike well-documented experimental systems, they are easily available, some even as a free Internet service. Some of these systems seem to follow very closely a *basic model* (a sort of advanced morphological transfer strategy), described in detail in this paper, which focuses on the translation from English to Spanish, and, in particular, on the mechanisms used by the systems to obtain the correct word order for the target language. The *basic model* is so simple that a laboratory assignment based on it allows students to discover interesting details about the operation of a number of real MT systems, such as the reordering rules used.

Keywords: teaching, commercial machine translation systems, word reordering.

1 Introduction

Many universities teach undergraduate and graduate courses dealing with the subject of machine translation (MT); part of these courses is expectedly devoted to teaching MT strategies. On the other hand, commercial MT systems are readily available, either as low-priced software packages for PCs (ranging from 30 to 300 euros) or as free Internet servers. A recent survey (Balkan et al. 1997) has dealt, among other aspects, with the use of commercial MT systems in teaching; while almost all respondents agree that "while using a working MT system to teach MT is definitely beneficial, it involves a huge amount of work". The authors "had hoped to find [...] that

[*] Supported by the Spanish Comisión Interministerial de Ciencia y Tecnología through grant TIC97-0941 and also by the Caja de Ahorros del Mediterráneo.

someone had done all the hard work" but got negative results. The recent increase in Internet availability of commercial MT systems may alleviate another problem described in the study, namely, that "[...] those interested in obtaining an MT system said they were prepared to invest very little" (a few hundred euros and from one day to one week). The work reported in this paper covers ways to use readily available commercial systems in teaching MT strategies. I do not claim to have done "all the hard work", but expect this proposal to be useful for the community of MT instrúctors.

In particular, a laboratory assignment is proposed, where the instructor puts forward a set of initial hypotheses about the strategy used by the systems and students work with test suites to find whether these hypotheses hold and to obtain relevant details about the particular rules used by the system. Laboratory work is designed for non-computer-science majors and has been tested during the last four years with third-year translation majors having only basic computer user skills (word processing, Internet navigation), although it may be used as well with computer science majors.

In the particular assignments presented here, the source language (SL) is English and the target language (TL) is Spanish, because most of our translation and computer science students are familiar with this pair.

The hypotheses are compatible with a transfer MT architecture (Arnold et al. 1994; Arnold 1993; Hutchins and Somers 1992) called here the *basic model* (see section 2). The basic model explains, at least partly, the behavior of the three systems covered: Globalink's *Power Translator Pro 5.0* or *Power Translator Deluxe* (PT), which are basically equivalent, Transparent Technologies' *TranscendRT*[1] (TRT), and Softissimo's *Reverso*[2].

A note of caution is necessary: the models proposed are derived from a black-box study of these systems, in the absence of documentation by the manufacturers; the models may therefore be partially incorrect or inaccurate, but this does not invalidate their use in the laboratory as long as they explain the basic behaviors observed and solve the question "Why do I get this *word salad*?". In particular, the models may be used to stress the mechanical or rule-based nature of machine translation in front of frequent student misconceptions about the behavior of computers, particularly among non-computer-science majors. Indeed, in one case, the rule set induced was completely confirmed by the contents of one of the system files (Mira i Gimènez and Forcada 1998). While some manufacturers claim that their products actually perform syntax analysis[3], close observation reveals behaviors that are not compatible with what most experts would call syntax analysis (for example, the ability to identify and correctly process simple constituents —such as noun phrases— regardless of their length).

2 The "basic model"

The basics of the behavior of the commercial MT systems examined may be conveniently explained using a *basic model*, a simple transfer architecture (Arnold et al. 1994; Arnold 1993; Hutchins and Somers 1992) which is more advanced than a morphological transfer system but cannot be properly called a syntactic transfer system because it does not perform full syntax analysis.[4] We have found this *basic model* to be simple enough for students to understand it and apply it to explain and predict the behavior of an MT system. The model is neither aimed at describing MT systems in general nor to encompass the basic problems of MT, but rather to describe in simple terms the behavior of a set of real, commercially available MT systems.

Four basic tasks are clearly distinguishable in the basic model: *morphological analysis*, which yields all possible lexical forms (LFs) for each surface form (SF) in the text[5]; *homograph*[6] disambiguation, which chooses one of the LFs, usually using simple rules based on the lexical categories assigned to neigboring words; The *transfer* task itself (see below); and *morphological generation*, which transforms each of the TL lexical forms (TLLFs) into the corresponding SFs.

The transfer task is organized around *patterns* representing fixed-length sequences of source-language LFs (SLLFs); two sequences are equivalent if they contain the same sequence of lexical categories. The system contains a catalog of the patterns it knows how to process. Patterns are not "phrases" or constituents in the syntactic sense, because they are flat and unstructured, but pattern detection is an advancement with respect to bare morphological analysis and may be considered as a rudimentary form of syntax analysis.

The *pattern detection* phase occurs as follows: if the transfer module starts to process the i-th SLLF of the text, l_i, it tries to match the sequence of SLLFs $l_i, l_{i+1}, \ldots$ with all of the patterns in its pattern catalog: the longest matching pattern is chosen, the matching sequence is processed (see below), and processing continues at SLLF l_{i+k}, where k is the length of the pattern just processed. If no pattern matches the sequence starting at SLLF l_i, it is translated as an isolated word and processing restarts at SLLF l_{i+1} (when no patterns are applicable, the systems resort to word-for-word translation). Note that each SLLF is processed only once: patterns do not overlap; hence, processing occurs left to right and in distinct "chunks".

Pattern processing takes the detected sequence of SLLFs and builds (using the bilingual dictionary) a sequence of TLLFs which may be completely reordered, with LFs added to it or deleted from it. The inflection information in TLLFs is generated so that agreement is observed inside the sequence if necessary. For instance, the English pattern article–adjective–noun (such as "the red tables") is turned into the Spanish sequence article–noun–adjective ("las mesas rojas"), after propagating the gender and number of "mesas" to both the article and the adjective.[7]

A finite catalog of fixed-length "frozen" sequences cannot possibly cover all of the possible forms a certain constituent (i.e., a noun phrase) may take, because of recursivity in grammar rules (for example, there is no theoretical limit to the number of noun phrases in the possessive case inside a given noun phrase). However comprehensive the catalog is[8], the system will always find unknown phrases: a pattern will match part of the constituent, process it as a complete constituent, and leave the trailing words up for further processing; this usually results in a "word salad", which would be very hard to interpret; however, having the basic model in mind, these garbled translations give invaluable cues as to which are the particular patterns in the system's catalog; this will be exploited in the laboratory assignment proposed.

3 Laboratory assignment

The purpose of this laboratory assignment is to discover the reordering patterns used by the three MT systems studied. It is designed for a two- or three-hour session (but may be cut down by reducing the number of machine translation programs studied), and requires substantial guidance by the course instructor.[9]

In the assignment, students will be asked to study the behavior of PT, TRT and Reverso when translating noun phrases of growing complexity from English to Spanish, to try to understand the strategy they use. Initially, they will study the translation of the 10 sentences represented by the expression I saw the [[senior] [computer] expert's] [large] desk. An acceptable translation, resulting from considerable reordering of words, is Vi el escritorio [grande] [del experto [de computadora] [mayor]].

Students will be told to assume that the system does not perform real syntactical analysis, but instead uses the strategy explained in this paper. For simplicity, they will be recommended not to consider articles as part of the patterns as a first approximation.

Students will write down, for each English sentence, the translation produced by each program, the nearest acceptable Spanish translation, and, where differences are significant, a possible explanation in terms of the proposed strategy. The following questions may be used to guide their work: "Can you identify parts of the sentence which have been independently processed? Which are the active patterns in each program? Why do we get incorrect translations for some sentences?"

If time allows, students will be invited to confirm the details of their hypothesis (patterns, etc.) with more noun phrases having the same sequences of lexical categories but different words (adjectives, nouns and nouns in the possessive case), or different composition. Make sure they do not introduce any homograph.

3.1 Hints for the instructor

What follows is an analysis of the results produced by each of the programs, to help the instructor guide the students during the assignment.

Power Translator Pro 5.0 (PT):

1. I saw the desk → Yo vi el escritorio: Acceptable. No reordering occurs.
2. I saw the large desk → Yo vi el escritorio grande: Acceptable. A reordering occurs in large desk, which may be explained with rule $R_1 : A\ N \rightarrow N\ A$, where N stands for a noun and A stands for an adjective.
3. I saw the expert's desk → Yo vi escritorio del experto: Acceptable except for articles. The reordering in expert's desk may be explained with a new rule: $R_2 : NG_1\ N_2 \rightarrow N_2\ \mathbf{d}\ N_1$ where NG stands for a noun in the possessive (genitive) case and **d** for the preposition de.
4. I saw the expert's large desk → Yo vi el escritorio grande de experto: Acceptable except for articles. The reordering in expert's large desk has to be explained with a new rule ($R_3 : NG_1\ A\ N_2 \rightarrow N_2\ A\ \mathbf{d}\ N_1$), because R_1 would have yielded *experto escritorio grande, and R_2 cannot be applied.
5. I saw the computer expert's desk → Yo vi el escritorio de experto de computadora: Acceptable except for articles. The reordering in computer expert's desk has to be explained with a new rule ($R_4 : N_1\ NG_2\ N_3 \rightarrow N_3\ \mathbf{d}\ N_2\ \mathbf{d}\ N_1$), because only R_2 may be applied and it would have yielded *computadora escritorio de experto.
6. I saw the computer expert's large desk → *Yo vi la computadora escritorio grande de experto. Unacceptable: PT splits the noun phrase computer expert's large desk. Only *expert's large desk* is reordered, using rule R_3, because PT has no rule matching the sequence $N\ NG\ A\ N$.
7. I saw the senior expert's desk → Yo vi el escritorio de experto mayor: Acceptable except for articles. The reordering in senior expert's desk, that is, in a sequence $A\ NG\ N$, has to be explained with a new rule, $R_5 : A\ NG_1\ N_2 \rightarrow N_2\ \mathbf{d}\ N_1\ A$, because only R_2 could have been applied, with the result *mayor escritorio de experto.
8. I saw the senior expert's large desk → *Yo vi el mayor escritorio grande de experto. Unacceptable: PT splits the noun phrase senior expert's large desk. Once again, only *expert's large desk* is reordered, using rule R_3. PT's catalog does not contain the pattern $A\ NG\ A\ N$.
9. I saw the senior computer expert's desk → *Yo vi la computadora mayor escritorio de experto. Unacceptable: PT splits the noun phrase senior expert's large desk. First, rule R_1 is applied to senior computer and then rule R_2 is applied to expert's desk. PT's catalog does not contain the pattern $A\ N\ NG\ N$.

10. I saw the senior computer expert's large desk → *Yo vi la computadora mayor escritorio grande de experto. Unacceptable: PT splits the noun phrase senior computer expert's large desk. First, rule R_1 is applied to senior computer, and then rule R_3 is applied to expert's large desk. PT's catalog does not contain the pattern $A\ N\ NG\ A\ N$.

Optional work with PT: As was explained in detail in (Mira i Gimènez and Forcada 1998), PT stores the patterns in an ASCII text file, named `engspan.pat` in directory `dicts`. An extended assignment may involve examining the file, or even modifying it to change PT's behavior.

TranscendRT (TRT): TRT's strategy is analogous to that of PT, but with some differences in rules: R_1 is applied to sentences #2, #6, and #10; R_2 is applied to sentences #3 and #6 (in #6, TRT applies the rule ignoring the possessive case in N_2); R_3 is used in #4; R_4 is used in sentence #5. The new rules are:

R'_5: $A\ NG_1\ N_2\ \rightarrow; N_2\ A\ \mathbf{d}\ N_1$ (the rule assumes that the adjective modifies the second noun; used in #7 and #10; in #10, TRT ignores the possessive case in N_2.).

R'_6: $A_1\ NG_1\ A_2\ N_2\ \rightarrow; N_2\ A_1\ A_2\ \mathbf{d}\ N_1$ (it reorders adjectives inadequately in this case but may be correct in other cases; used in #8).

R'_7: $A\ N_1\ NG_2\ N_3\ \rightarrow N_3\ A\ \mathbf{d}\ N_2\ \mathbf{d}\ N_1$ (it assumes that the adjective affects the third noun; used in #9).

In addition, TRT deletes the pronoun *Yo* before the verb *vi*, translates *expert* as *perito* instead of *experto* and has a different treatment for articles.

Reverso: Reverso's strategy is very similar, with some differences in the rules applied. The main difference is the addition of the Spanish preposition *a*, mandatory before direct-object noun phrases having a person as their head. Rules R_1–R_4 are as in TRT, and R_5 as in PT. The new rules are:

R''_6: New rule: $A_1\ NG_1\ A_2\ N_2\ \rightarrow\ N_2\ A_2\ \mathbf{d}\ N_1\ A_1$, used in #8.

R''_7: New rule: $A_1\ N_1\ NG_2\ A_2\ N_3\ \rightarrow\ N_3\ A_2\ \mathbf{d}\ N_2\ \mathbf{d}\ N_1\ A_1$, used in #9.

4 Concluding remarks

I have shown how a very simple model of MT may be very useful to obtain a rather detailed explanation —including the formulation of rules— for the word-reordering behavior of three readily available commercial MT systems, and how this work may be organized as a laboratory assignment in which the students use the model to formulate the particular rules used by each system under the guidance of the laboratory instructor.

Notes

[1] http://www.freetranslation.com

[2] http://proto.softissimo.com/reverso/asp/textonly/default.asp

[3] "**Analyze syntax:** TranscendRT determines the function of each word in the sentence" (http://www.transparentlanguage.com/ets/about/transcendrt.htm).

[4] The *basic model* may also be considered as a special case of what Arnold et al. (Arnold 1993, sec. 4.2) call a *transformer architecture* or what Hutchins and Somers (Hutchins and Somers 1992, sec. 4.2) refer to as a *direct* architecture.

[5] In particular, morphological analysis takes each SF or word (e.g., *taught*) and builds one or more LFs per word, consisting of a lemma or canonical form (*teach*), the lexical category (*verb*), and inflection information (*past tense*)

[6] A homograph is a SF having more than one LF.

[7] In addition, the transfer module may maintain "state" information to ensure left-to-right interpattern relationships such as subject–verb number agreement. State information may be updated after each pattern is processed.

[8] Consider also that the number of patterns grows dramatically with length.

[9] This assignment should be placed after a preliminary assignment in which a comparison of the translation of a set of sentences proposed by the instructor with the translation of each word in isolation —one word per line, with a blank line between words— reveals processes such as context-dependent homograph disambiguation (part-of-speech tagging), use of multiword units, *word reordering*, and agreement enforcement.

References

Arnold, D. (1993). Sur la conception du transfert. In Bouillon, P. and Clas, A., editors, *La traductique*, pages 64–76. Presses Univ. Montréal, Montréal.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., and Sadler, L. (1994). *Machine Translation: An Introductory Guide.* NCC Blackwell, Oxford.

Balkan, L., Arnold, D., and Sadler, L. (1997). Tools and techniques for machine translation teaching: A survey. Technical report, University of Essex, Colchester, Essex, U.K. URL: `http://clwww.essex.ac.uk/group/projects/MTforTeaching`.

Hutchins, W. and Somers, H. (1992). *An Introduction to Machine Translation.* Academic Press.

Mira i Gimènez, M. and Forcada, M. L. (1998). Understanding PC-based machine translation systems for evaluation, teaching and reverse engineering: the treatment of noun phrases in Power Translator. *Machine Translation Review (British Computer Society)*, 7:20–27. (available at `http://www.dlsi.ua.es/\~{}mlf/mtr98.ps.Z`).