Discovering Machine Translation Strategies Beyond Word-for-Word Translation: a Laboratory Assignment

Juan Antonio Pérez-Ortiz
Mikel L. Forcada

Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain

Abstract

It is a common mispreconception to say that machine translation programs translate word-for-word, but real systems follow strategies which are much more complex. This paper proposes a laboratory assignment to study the way in which some commercial machine translation programs translate whole sentences and how the translation differs from a word-for-word translation. Students are expected to infer some of these extra strategies by observing the outcome of real systems when translating a set of sentences designed on purpose. The assignment also makes students aware of the difficulty of constructing such programs while bringing some technological light into the apparent "magic" of machine translation.

1 Introduction

A common misconception among students, specially those not familiarized with machine translation (MT), is that MT systems follow a strategy similar to that implemented in early MT programs in the 50's. This strategy, usually known as word-for-word translation[1], ignores inter-word dependencies considering each word in a sentence in isolation, and lacks any kinds of intermediate representations.

Obviously, this kind of strategies produce very poor results, even when the source language (SL) and the target language (TL) share similar lexical, morphological, syntactical and semantical structures. In fact, this basic approach to MT is what we might expect if we asked a non-expert to design a MT system. The outcome would be comparable to that obtained from "someone with a very cheap bilingual dictionary and only the most rudimentary knowledge of the grammar of the target language: frequent mistranslations at the lexical level and largely inappropiate syntax structures which mirrored too closely those of the source language" (Hutchins and Somers 1992, p. 72).

On the one hand, current real MT programs implement techniques much more advanced than word-for-word translation. Although there are a lot of situations in which they still keep on generating wrong translations, MT systems perform a deep analysis on sentence as a whole, implementing processes such as context-dependent homograph[2] resolution, special processing of multiword units (such as idioms), word reordering, agreement enforcement, or exception handling.

Nowadays, on the other hand, commercial systems whose translations may be considered acceptable to some level are available at low or medium prices, or even freely on the Internet; they have become an affordable tool for helping the task of the machine translation instructor.

Our proposal is a laboratory assignment where students discover some of the multiple processes which go beyond a simple word-for-word strategy and are implemented in real MT systems, and how they are better than the word-for-word approach. Laboratory work is mainly designed for non-computer-science majors but it may be used as well with computer-science majors. The source language (SL) is English and the target language (TL) is Spanish. It has been succesfully tested for six years with third-year translation majors with very basic computer skills in general.

Machine translation majors learn also the advantages and disadvantages of using MT programs: these programs are enormously imperfect but they still can be useful. Furthermore, the assignment may help non-computer students to give up some misconceptions (sometimes a complete ignorance) about the algorithmic behaviour of computers.

2 Word-for-word and Indirect Architectures

A word-for-word translation strategy can be described as a three-phase process (Hutchins and Somers 1992, p. 72):

a) The first phase consists of a rudimentary morphological analysis where each superficial form (SF) in the SL is converted into its corresponding lexical form (LF). Homograph disambiguation is not implemented in this approach.

[1] Hutchins and Somers (1992) call it *direct* translation whereas Arnold (1993) calls it a *transformer architecture*.

[2] A *homograph* is a superficial form (SF) having more than one lexical form (LF). The SF is the form in which a word appears in a text (for instance, *rang*); the LF is composed of a lemma (*ring*), a lexical category (*verb*) and inflection information (*past tense*).

b) A bilingual dictionary is looked up in the second phase in order to translate each LF to its corresponding LF in the TL.

c) Finally, the LF in the TL is inflected to obtain the translation (some local reorderings are probably done in this phase as well).

There is an even more radical approach in which no analysis or reordering takes place: each superficial form is looked up in a dictionary, leading directly to a word in the target language. This is by far the most common preconception among students.

Although the assignment is designed for students to discover new rules beyond word-for-word translation, no particular model needs to be assumed initially by them. Anyway, we have found that having in mind a basic transfer architecture, which is more advanced than a morphological transfer but simpler than a syntactic one, is adequate for students to bring some light into the observed behaviour of the MT programs when translating whole sentences.

3 Laboratory Assignment

The purpose of the assignment is to discover which processes beyond the basic word-for-word approach are performed by MT systems. To avoid wrong generalization, three different programs are covered: Globalink's (now Lernout and Hauspie) Power Translator Pro 5.0 (PT)[3], Transparent Technologies' TranscendRT (TRT)[4] and Softissimo's Reverso[5].

Students compare the translations given by the MT systems for a set of sentences with the translation of each word in isolation. To obtain the latter, sentences must be typed one word per line with an additional blank line between words.[6]

The assignment is not aimed to infer the exact rules which are aplied by the programs, but to find where these rules act and to set some hypotheses about their behaviour. It is dessigned for a two-hour session (but may be cut down by reducing the number of machine translation programs studied) and requires substantial guidance by the course instructor.[7]

Students are given five sentences in English and are told to write down the translation given by each program to the isolated words of each sentence, the output given when considering sentences as a whole, and the nearest acceptable Spanish translation . Then they are suggested to formulate detailed hypotheses about the observed differences.

The five sentences are chosen so as they are correctly translated by at least one of the programs. Students' work may be guided by the instructor by formulating some questions, such as,

a) why do you think that in some cases the translation obtained for a simple word is completely different than the translation of the same word when being part of a whole sentence?

b) which rules do you think the program uses for choosing one translation or the other one?

c) apart from deciding which translation to choose for each word, which other operations does the program perform? would you dare advancing an explanation of the translation strategy implemented by the program?

In some cases, the preceding questions may be answered by looking up the program's dictionary, when available; students are strongly recommended to do so. Access to multiword and single word dictionaries is explained by the instructor in the beginning of the assignment.

If time allows, students are invited to repeat the previous steps with new sentences they propose in order to confirm some of the formulated hypotheses or infer new ones.

4 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.[8]

Power Translator 5.0 (PT). As an example, we analyse the results of PT in detail.

1. My tailor is rich. Word-for-word: Mi / adapte / es / rico. Whole sentence: Mi sastre es rico.

   The word tailor is a homograph in English. PT considers it as a verb (imperative of to tailor) when seen in isolation, giving adapte in Spanish; on the other hand, the second translation gives sastre (a noun). PT may have considered the preceding word (my) and discarded the unusual combination of possesive adjective and verb, choosing the more usual combination of possesive adjective plus noun instead. The disambiguation rules are correct in this case.

   A deeper discussion may be done as well on the possible influence of the adjacent word is.

2. Artificial intelligence systems can think. Word-for-word: Artificial / la inteligencia / los sistemas / poder / piense. Whole sentence: Los sistemas de inteligencia artificial pueden pensar.

   The more obvious transformation in the second translation is the reordering done on the elements of the noun phrase artificial intelligence systems, which produces the right Spanish form los sistemas de inteligencia artificial. Moreover, PT inserts the definite article los (the in English) before the noun phrase, following the correct criterion in Spanish (the insertion of articles is also observed in the word-for-word translation where both nouns, namely intelligence and systems, are preceded by a corresponding definite article). It may

[3] http://www.lhsl.com

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

[5] http://www.reverso.net

[6] After obtaining the word-for-word translation, these extra lines can be removed to obtain the translation of sentences as a whole.

[7] In our course, this assignment is followed by another one in which students infer the reordering rules applied by MT systems to growing-complexity noun phrases (Forcada 2000).

[8] Students are suggested to number the sentences in order to identify easily the corresponding translation.

be considered as well that a homograph resolution is taking place with the word can (possibly a noun) and that PT is inserting new words, such as the preposition de.

In this case, when translating a verb such as can in isolation, the program does not choose the imperative as before but the infinitive form (poder). It seems that PT is capable of distinguishing between lexical and modal verbs. Furthermore, when considering the whole sentence, the adequate inflected form of the verb (pueden) is used; this is probably done by determining the number of the preceding noun phrase, a task easy in English since it suffices with determining the number of the last word in the noun phrase. PT deals, therefore, with agreement between verb and subject.

The word think in isolation produces again the imperative piense in Spanish. In the second translation, we may infer that the system is detecting the presence of a modal verb (can) before think and using, as a result, the infinitive form (pensar) in Spanish.

3. Machine translation programs cannot translate complex texts. Word-for-word: Elabore / la traducción / programa / no poder / traduzca / el complejo / los textos. Whole sentence: Los programas de traducción automática no pueden traducir textos complejos.

   Some features already discussed appear: homograph resolution (machine and programs can be a noun or a verb, complejo can be a noun or an adjective), article insertion (los programas), agreement propagation (no pueden), detection of modal verbs and reordering of two noun phrases. There are, on the other hand, new features whose discussion follows.

   The most significant aspect is the translation of machine translation as the right traducción automática instead of the expected traducción de máquina. Looking up machine in the PT's dictionary reveals that it can be a noun (máquina) or a verb (elaborar); therefore, where does automática come from? The answer is in the multiword dictionary which makes PT treat machine translation as a single unit (the same way as, for instance, machine gun is translated as ametralladora and not as pistola de máquina).

   The students should also observe the correct agreement between the words textos and complejos (masculine and plural), and the absence of an article before this noun phrase; it might be infered that the addition of a definite article takes place before the subject but not before the direct object,[9] where the article is less usual in Spanish.

4. The computer expert's desk is large. Word-for-word: El / la computadora / el experto / el escritorio / es / grande. Whole sentence: El escritorio de experto de computadora es grande.

   Reordering rules make it possible to attain an acceptable translation, although experto en computadoras would be a better choice than experto de computadoras. If we replace expert by technician, the correct translation técnico en computadoras is directly obtained since the multiword dictionary includes an entry for it. Students can add the unit computer expert to the dictionary and reobtain the translation of the original sentence.

5. The computer expert's desk is full. Word-for-word: El / la computadora / el experto / el escritorio / es / lleno. Whole sentence: El escritorio de experto de computadora está lleno.

   This sentence is very similar to sentence 4; the only difference is the adjective full instead of large. PT chooses the right translation in Spanish for the verb is depending on the adjective: es in sentence 4 and está here. The adjective full is handled as an exception; this can be corroborated by looking up both English adjectives in the dictionary: the adjective full contains a reference to a *parset* (paradigm set) which may be shown to handle this exception by adding adjectives to the dictionary.

TranscendRT (TRT). Basically, the main difference in word-for-word translation between TRT and PT is the preference of noun over verb when disambiguating homographs. For whole sentences, the reordering rules are different, and some extra articles are inserted where appropiate. TRT also considers the adequate form of the verb to be in Spanish (ser or estar) when dealing with the adjectives large or full.

Reverso. Reverso's translation of these sentences is very similar to PT's; it also handles traducción automática as a special unit and reorders correctly the noun phrases. Reverso does not distinguish between the adjectives large and full and inserts definite articles before the noun experto.

5 Concluding Remmarks

This paper proposes a laboratory assignment which may be useful to help students abandon the misconception that MT programs are very simple word-for-word translation engines, and infer from samples the particular rules implemented by these programs. The assignment may be useful as well in helping MT students to develop a suite of sentences in order to evaluate MT programs (Trujillo 1999, chapter 10).

Acknowledgements

This work has been supported by the Spanish Comisión Interministerial de Ciencia y Tecnología through grant TIC97-0941 and by the Caja de Ahorros del Mediterráneo.

References

Arnold, D. (1993). "Sur la conception du transfert". In P. Bouillon and A. Clas (eds.), *La traductique*, (pp. 64–76). Presses Univ. Montral, Montral.

[9] If PT is considered less complex than a syntactical transfer system, the program should follow a heuristic rule to identify the subject and the direct object. The simplest explanation is that subject precedes the verb and direct object follows it.

Forcada, Mikel L. (2000). "Learning machine translation strategies using commercial systems: discovering word-reordering rules". In *MT2000: Machine Translation and Multilingual Applications in the New Millenium*.

Hutchins, W. John and Harold L. Somers (1992). *An introduction to machine translation*. Academic Press, London.

Trujillo, Arturo (1999). *Translation engines: techniques for machine translation*. Applied Computing. Springer, London.