Abstract
In this paper, I look at some of the problems of Machine Translation (MT) as compared to multi-lingual text processing and text retrieval. In particular, I discuss the question of the evaluation of the output of MT systems, as opposed to the output of other NLP applications. Finally I raise the question of the political implications of the choice of input and output languages for MT systems within the larger context of a world where electronic communication renders access to textual information crucial for the maintenance of less established languages.

Machine Translation
It is well-known that Machine Translation (MT) is a type of application which presents most of the most difficult tasks of Natural Language Processing (NLP). It involves at the very least the following three components:

a. analysis of the source text,

b. generation of the target text, and

c. mapping between the two representations needed for tasks (a) and (b)

The choice of levels of representation for (a) and (b) as well as the type of mapping—interlingua, transfer, or direct approach— for (c) are the basic decisions which have to be made when designing an MT system.

These choices depend on a number of factors, including:

- the number of language pairs involved;
- the goal of translation (information or dissemination);
- the type of texts;
- the desired extensibility of the system;

MT systems based on banks of previous translations, or based on statistical predictions, still have to choose at what level to recognize the translation units on which they operate, and how to perform the mapping. However, I will not discuss that type of system in more detail.

Extension of MT systems
Two of the main difficulties encountered in real-world MT (as opposed to research systems and prototype design) are 1) the extent to which a given system can be modified by end-users, and 2) the extent to which a system can be modified to deal with new language pairs.

1. The most general case is the extension of existing systems to treat new domains. The simplest way is to modify the user dictionaries. This includes:

- adding and removing lexical entries,
- altering the specifications of meaning, subcategorization or co-occurrence requirements of a number of items.

One of the problems here is to define a core part of the lexicon, which should not be modified by users. Another is to devise a user-friendly interface which allows non-linguistically trained users to make the required modifications in accordance with the way the system actually uses that kind of information. While the former problem is present to a certain extent in all NLP systems, the latter is particularly acute when the users are expected to provide such information for lexical entries in a language of which they are not native speakers. Typically, a translator using an MT system will be a native speaker of the source or target language, but not of both.

2. The more challenging case of extension of MT system is the adaptation of existing systems to create new language pairs. There are several options:
Rewrite the input and/or the output grammars.

Modify existing grammar to treat a related language (e.g. WCC used the German grammar as the basis for a new Swedish grammar; in the prototype of the ELU system at ISSCO, the first Italian grammar was based on a previous French grammar).

Create a new transfer component to make a new language pair out of already existing language pairs (e.g. the SLT project: English-Swedish, and English-French → French-Swedish). This can be done manually, or automatically (Rayner et al. 1996).

Reusability of grammars and of analyses

For maximum reusability of grammars, each source language grammar in the system must be as independent of the target language grammar as possible. This is one of the main arguments against systems based on direct transfer between two languages (this argument about the reusability of grammars is orthogonal to the debate between about the choice of interlingua as opposed to transfer, as long as transfer is not direct). Ultimately, the ideal solution is to have fully reversible grammars, which can be used both as source and target grammars (Estival 1994).

Differences between MT systems and Multi-lingual systems

The most obvious difference is that while there must be a text generation component in an MT system, most other multi-lingual systems do not require textual output of the same type as their input.

However, other systems may possibly require a deeper “understanding” of the texts they treat. It is well-known that MT can, in a number of cases, get away with keeping ambiguities from the source text (although it is not a strategy to be recommended), while Information Retrieval (IR) can’t, and neither can Text Retrieval (TR)

On the other hand, there are problems specific to multi-lingual systems, which are assumed to be solved in MT:

- language recognition (although this is one of the tasks of a speech-to-speech translation system such as Verbmobil (Wahlster 1993));
- choice of an encoding scheme for writing systems which require more than pure ASCII (diacritics; or more difficult, syllabaries or ideograms rather than alphabets);

Evaluation of NLP Systems

While the evaluation of NLP systems in general has become an extremely important area of research in the past decade and great progress has been made in the design of strategies for assessing the quality of the output of a number of types of NLP applications (Sparck Jones & Galliers 1996), much less progress has been made for MT.

In fact, it can be claimed that the most challenging aspect of MT, as compared to other NLP systems, is the problem of the evaluation of the output of an MT system (King & Falkedal 1990).

Indeed, the task is not specifically an NLP task, but involves a much more fundamental question: “How do you evaluate the quality of a translation?”.

The two main ways of evaluating NLP systems have been the use of test-suites and the use of benchmark corpora. Both have been used extensively to evaluate MT systems, with very mixed results.

The goal of the European LRE project TSNLP was to design a methodology for the use of test-suites for NLP. It is interesting to note that on the one hand, it found that most existing test-suites had been created for testing and evaluating MT systems (Estival et al. 1994), but that on the other hand, it decided not to include MT as one of the applications for which test-data was created in its multi-lingual database of test material (see Lehmann et al. 1996 for a description of that database). The main reason for not including MT was the difficulty of specifying the expected output for a given input sentence.

While the advantage of test-suites is that they provide a systematic and complete way of testing applications, the advantage of corpora (e.g. the set of newspaper articles for the ADAcel project MEMTAH (Indonesian → English)) is that they present the system with the type of natural data it is likely to encounter. However, there is still the problem of specifying what the expected output should be.

The political implications of language choice in NLP

Which languages are accessible through technology has always been a very important political as well as cultural issue (what language is chosen as the “national” or “official” language of the government, the media and for education, and thus gets to be printed and to have dictionaries), but this issue is rapidly becoming even more crucial in areas and in countries where the everyday language, which may be printed, is not yet available on computer systems and for electronic communication (see Omoniyi 1996).
This issue is now connected to the problems of survival for many non-dominant languages, i.e. not English, Japanese and a few of the major European languages, but most of the world languages. The Pacific Rim area is a good example of a multi-cultural and multi-lingal community, where a large majority of the languages actually used today are in danger of disappearing in the next few decades. Languages which will not be available for electronic communication can be considered doomed to a more rapid extinction than could be foreseen even a few years ago.

To ensure the survival of a language, the first step is the creation of an encoding scheme for the writing system of the language. This may involve choosing among alternative local systems, and must result in compatibility with international standards.

The next step is the development of lexical resources: compiling terminological lists and writing dictionaries (both monolingual and bilingual); and collecting written texts into electronically available corpora (such as those available through the Linguistic Data Consortium or the ACL/ECI initiative).

Finally, it must be noted that, in MT, the choice of direction of translation has not been innocent either: many more systems translate from English than into English.

References

Estival, Dominique, Kirsten Falkedal et al. (1994). Analysis of Existing Test Suites, Report to LRE 62-089 (D-WP1), University of Essex.


