NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS:
The copyright law of the United States (title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material. Any copying of this document without permission of its author may be prohibited by law.

CMU-CS-84-142

The Design Philosophy of Personal Machine Translation System

Masaru Tomita
Computer Science Department
Carnegie-Mellon University
Pittsburgh, PA 15213, USA
1984

Abstract

Most existing practical Machine Translation (MT) systems are designed to translate documentation, such as technical papers and manuals. However, there is a growing need for translating not only large texts such as these, but also personal short texts such as letters and informal messages. The conventional MT systems, which are intended to translate large texts, are not very suitable for these kinds of small jobs. We need a different type of system for these which, throughout this paper, we refer to as a Machine Interpretation (MI) system, in contrast with the conventional Machine Translation (MT) systems. This paper contrasts the design philosophy of an MI system with that of an MT system. We see that the total amount of human assistance is crucial in MT systems, but not necessarily crucial in MT systems; on the other hand ease to use is crucial in MI systems, but not necessarily crucial in MT systems. This paper then discusses what an MI system should look like under these conditions.

Keywords

Machine Translation, Natural Language Processing, Man-Machine Interaction, Human-assisted Machine Translation.

Acknowledgements

I would like to thank Herbert A. Simon and Jaime Carbonell for discussions that led to many useful ideas, and Cynthia Hibbard for helping to produce this document.

This research was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory Under Contract F33615-81-K-1539. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

Table of Contents

.1 Introduction

Most existing practical Machine Translation (MT) systems are designed to translate documentation, such as technical papers and manuals. However, there is a growing need for translating not only large texts such as these, but also personal short texts such as letters and informal messages. The conventional MT systems, which are intended to translate large texts, are not very suitable for these kinds of small jobs, as we see later. We need a different type of system for these which, throughout this paper, we refer to as a Machine Interpretation (MI) system, in contrast with the conventional Machine Translation (MT) systems. The MI system, however, should not be just a miniature version or extended version of the conventional MT systems. We must approach the MI systems with a totally different design philosophy.

This paper contrasts the design philosophy of an MI system with that of an MT system. This paper then discusses what an MI system should look like under these conditions.

2 MT systems vs. MI systems

In the 1950's, Machine Translation projects were started to reduce the cost of translating technical documents by replacing human translators by computers. Since then, the goal of the MT systems has not been changed much. In fact, the goal of existing practical MT systems is to reduce the cost of human translators.

In 1960, Bar-Hillel argued that fully automatic high quality translation was not only practically, but also theoretically impossible without a full understanding of a text [1]. Therefore, all practical MT systems, without exception, are not fully automatic, but require human assistance in some way. The methods of human assistance can be classified into four categories:

- (1)**Rejection**, Sentences which the system cannot handle are rejected. Only tractable sentences are translated by the system. The rejected sentences are then translated by human translators.
- (2)**Pre-editing**, Source texts are edited by humans to make them fit the syntax and the vocabulary the system can handle.
- (3)**Post-editing**, The system takes unedited source texts and outputs target texts which require substantial human post-editing.
- (4)**Interactive method**, The system requires neither pre-editing nor post-editing; instead, it requires interactive human assistance during the translation.

To make an MT system practically useful, the cost of the assistance a human gives to the system must be less than the cost of providing a human to translate the whole text without a computer. Consider the case where one has a document which a human translator would translate in 30 days without the system. If 60 days of human assistance are required for the system to do the reasonable translation, then the system would be certainly too costly to be useful. If the MT system requires 30 days of human assistance, the system is probably still not workable. This cost of human assistance is crucial for developing practically useful MT systems.

Next let us consider an MI system, that is intended to translate personal and small texts. The goal of the MI system is not to reduce the cost of translating a document. Rather, the goal is to enable the user to translate a small text without a human translator or specialist, who are not available to translate one or two paragraph texts immediately on demand. Because an MI system has a goal different from that of the MT systems, the conditions mentioned in the previous paragraph are not necessarily required for an MI system. In other words, the cost of human assistance is not necessarily crucial in an MI system. Suppose a small text can be translated in 3 minutes by a human translator. If an MI system requires 6 minutes of non-specialist human assistance to do the job, the user probably does the job by himself using the MI system, rather than calling a human translator. This is because one does not want to and may not be able to call a human translator each time one has such a small job. Thus, it is more acceptable for an MI system to count on human assistance than it is for a large scale MT system, because the MI system's jobs are always small.

On the other hand, in order to make an MI system practically workable, the system must satisfy a number of conditions that MT systems need not satisfy. First of all, in an MI system, the type of assistance required must be knowledge that all users prossess, not expert knowledge requiring specialists. The system should assume that a user speaks only his own language (the "source"), and does not know anything about a foreign language (the "target"). Also, it should be assumed that the user is neither a computer engineer nor a linguist. Thus, the user must not be required to know any foreign language, computer science or linguistics. This is because the user does not want to call any of these specialists each time he has a small job. On the contrary, an MT system can afford such specialists, as long as the total cost of human assistance is less than the cost of having human translators to do the same job without the MT system. Indeed, most existing practical MT systems require specially trained persons.

A second condition for an MI system to be practically workable is that the system response must be reasonably quick since an MI system runs on an interactive, real-time system. The user does not want to wait for minutes each time he translates a small text. By contrast, since an MT system usually runs

as a batch system, the response time is not such an important factor; only the time for human assistance matters.

The final condition for an MI system is that the system must be reasonably inexpensive so that every user can afford to run it in his home or office. MT systems are usually very expensive and installed only at major institutions. The user cannot bring a text to an institution each time he has a small job. Personal computers are therefore ideal for running MI systems.

To summarize, an MI system has the following tolerance. It can count on human assistance more than an MT system. However, it has 3 major constraints. It must be easy to use, respond reasonably quickly, and be affordable for every user.

3 Design Decisions

Considering this tolerance and the three constraints of an MI system, we then discuss the four main design decisions we might make.

3.1 Translation Direction

Assuming that the user knows only his own language, we can think of two translation directions; 1) foreign language to user's language, and 2) user's language to foreign language. If the system is to translate a document, then Direction 1 is probably more important than the other. However, for personal texts, neither of the directions is more important than the other. But we shall claim that Direction 2 is more feasible, because Direction 1 has a major problem; how to provide input. Although output can be generated by the system automatically, input must be provided by the user himself. For example, there is no easy way for English speakers to input Japanese sentences. Thus, in what follows, we focus on an MI system that translates the user's language into a foreign language.

3.2 Human Assistance Method

It is fairly easy to show that the interactive method is the most suitable for an MI system that translates the user's language into a foreign language. The other three methods of human assistance are inappropriate in the following ways: Rejection eventually requires a human translator for rejected sentences. To pre-edit an input text, the user must know the exact grammar and vocabulary the system can handle, and we have assumed that the user is neither a computer scientist nor linguist. Finally, to post-edit an output text, the user must know the foreign language.

In the interactive method, the user inputs a sentence. The system may then ask the user several questions, and on receiving answers, the system finally outputs a foreign sentence. Questions are asked to resolve grammatical ambiguity, lexical ambiguity, pronoun reference, unknown words, and mis-typing. The questions must be asked in the user's language, and must not require any knowledge of the foreign language, computer science or linguistics to answer them.

3.3 How Much Semantics

How much semantics should be included in an MI system is difficult to determine. Here is the tradeoff: The more semantics are included, the less ambiguity will be found and the fewer questions the system will ask the user, but the more expensive and the less affordable the system will be. We should first build a system with very little semantics, see how many questions the system will ask, and then think about reducing the number of questions with more semantics.

3.4 Size of Dictionary

The bigger vocabulary the system has, the fewer words of input will be unknown and therefore the fewer questions the system will ask. However, the bigger the vocabulary, the more memory will be needed and the slower the system will be. The system should not have a dictionary as big as the conventional MT systems . A reasonable size for the vocabulary may be around 10,000 words.

4 Concluding Remarks

In this paper, we have introduced the notion of a Machine Interpretation system. The system goal and the design philosophy of an MI system are different from those of the conventional MT systems. We saw that the total amount of human assistance is crucial for MT systems, but not necessarily crucial for MI systems; on the other hand, ease to use is crucial in MI systems, but not necessarily crucial in MT systems.

An experimental English-to-Japanese MI system has been built by the author's group, and is being implemented at Kyoto University in Japan [3]. It embodies the design philosophy discussed above and takes English as the source language and Japanese as the target language. Thus, the system translates an English sentence into Japanese, asking the user several questions in English interactively. This MI system, however, was realized by modifying an MT system [2] which is intended to translate technical texts. Therefore, the MI system is also limited to translating technical texts.

An MI system for wider domain is now being developed at Computer Science Department, Carnegie-Mellon University. Thus far, only parts of the system have been implemented in MACLISP on TOPS-20 to demonstrate the feasibility of our approach to specific problems [4] [5].

References

[1] Bar-Hillel, Y.
A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation.
Addison-Wesley, Massachusetts, 1964; .

[2] Nishida, T. and Doshita, S.
An Application of Montague Grammar to English-Japanese Machine Translation.
Proceedings of conference on Applied Natural Language Processing :156-165, 1983.

[3] Tomita, M., Nishida, T. and Doshita, S.
An Interactive English-Japanese Machine Translation System.
Forthcoming (in Japanese), 1984.

[4] Tomita, M.
Disambiguating Grammatically Ambiguous Sentences by Asking.
In *Proceedings of the 10-th International Conference on Computational Linguistics.* 1984.

[5] Tomita, M.
LR Parsers For Natural Languages.
In *Proceedings of the 10-th International Conference on Computational Linguistics.* 1984.