Machine Translation Systems in Practical Use

Martin Volk, Mirjam Oberholzer
University of Zurich,
Department of Computer Science, Computational Linguistics
Winterthurerstr. 190, CH-8057 Zurich
volk@ifi.unizh.ch, moberh@cl.unizh.ch

Abstract

The following paper summarizes part of the findings that resulted from a seminar on Machine Translation in Practical Use, which was held in September 1996 at the University of Zurich. In the seminar, fundamental decision-making criteria regarding introduction and choice of an MT system were discussed. The focus of this paper is on currently available translation tools for German-English translation, especially MT systems, and their usefulness in a business environment. To this purpose, we have concentrated on a survey of commercial MT systems and the criteria for their use. An outlook at possible developments in MT concludes the paper.

1 Introduction

Research in machine translation (MT) goes back to the early days of computer science (cp. [Hut86]). Over time it has been realized that fully automatic machine translation of arbitrary natural language material is a very complex if not impossible task. Therefore, the focus shifted from automatic machine translation to human aided machine translation and lately to machine aided human translation. The computer is now understood as a powerful tool that can support the human translator and increase her productivity. Some companies report on a 50% productivity gain after careful introduction of appropriate translation tools.

In this paper we will first give a brief survey of translation tools including machine translation systems. We will then describe some commercial machine translation systems that translate between English and German. In section 3 we will discuss criteria for the employment of machine translation in a company or administration unit. We conclude by looking at some of the future directions of machine translation.

2 Translation Tools

In a rough approximation, the translation activity of a human translator consists of understanding a sentence in the source text, finding the correct and appropriate words in the target language, and producing the target language sentence. A closer look reveals that this activity involves a multitude of smaller steps. Understanding a sentence in the source text may involve the lookup of technical words. Finding the corresponding words in the target language will also include looking up words in special bilingual terminology databases. The decision for a particular translation of a technical term has to be made according to industry and company standards and has to be kept consistent throughout a text. For the translation of some constructions it may be necessary to consult a grammar book. The target text needs to be checked for correct spelling and, finally, it has to be formatted like the source text. The computer can help in many aspects of this activity.

2.1 A Survey of Translation Tools

Translation tools can be classified in order of increasing complexity.

1. The most general tool for every translator must be a state-of-the-art **word processing** system. This should include a language module for the target language including keyboard layout, hyphenation, spelling checker and thesaurus. The keyboard layout ensures that all target language characters are easily accessible. The hyphenation and spelling modules help in checking the correctness of the target language text and the thesaurus is important to broaden the translator's active vocabulary.

   In addition, a spelling checker for the source language may help in preparing the source text for machine translation.

2. The word-processor should be supplemented by **online lexical and grammatical resources**. Currently this means mostly monolingual or bilingual dictionaries on CD-ROM. These can be interfaced to the word-processor in such a way that any word can be looked up with a mouse click. Even some grammar books are available on CD-ROM (sometimes hooked up to the dictionaries) and thus allow for faster access.

3. In addition to general lexical resources and thesauri, professional translators need **terminology databases**. These are collections of technical terms and their definitions and translations. These collections are usually subdivided into subject areas such as automotive parts, computer hardware, anorganic chemistry and so on. Some terminology collections may be purchased but others have to be collected over time. Any terminology database must allow for the modification and extension of its entries.

4. Often it is necessary to translate a document that is only a slightly modified version of an already translated document (e.g. multiple versions of a software manual). In this case a **translation memory system** can be very helpful. Such a system stores sentences together with their translation. When another sentence is to be translated, the system searches its translation memory to find if the sentence or a similar sentence has been translated in the past. The degree of similarity can be determined by the user. A translation memory system can also be advantageous if documents contain many fixed modules as is often the case with business letters.

5. Finally, there are **translation systems** that automatically or interactively translate a text. Translation systems are complex programs that come with huge lexicons and grammar modules. The more elaborate systems even include semantic features (i.e. they differentiate between animate and inanimate objects, time and place names). Translation systems work best when the source text conforms to

some principles of simplicity as we will discuss below. Nevertheless it is necessary to correct the output of a machine translation system if the text is meant to be published. This type of correction is called 'postediting'. It stands in contrast to tailoring the text before the translation starts, which accordingly is called 'preediting'.

In a recent EU project it has been investigated in some depth how these translation tools can be integrated into a translator's workbench (cp. [KAT95]). The authors come to the conclusion that "... most companies aim at particular tools only, and that there is presently no single company offering tools for all aspects of electronic writing."

2.2 Some Commercial Machine Translation Systems

For some time, commercial machine translation systems have been either affordable and of low quality or they were satisfactory but very expensive. In the last two years this has changed at least for the language pair English - German. There are still at least two high price systems (Logos and Systran), but there are now also six systems in the price range of less than US $2000. They all run on PCs under Windows. We have recently had the opportunity to install and evaluate these systems. The evaluation included tests on lexical and grammatical coverage as well as experiments to measure usability and translation quality (cp. the article by Amgarten and Merz in this volume).

German Assistant is a translation system by Globalink hooked to the multilingual word-processor Accent Duo. Switching from word-processor to translation system and vice versa is awkward. The translation quality is rather low compared to the other systems. As a special feature, German Assistant allows interactive translation where the user may choose among several translations of a word.

Langenscheidts T1 is based on the METAL system, a large and expensive system marketed by Siemens for some years. T1 has been ported to PCs by a company called GMS and is marketed by Langenscheidt Verlag. Among the systems, it ranks first in translation quality while maintaining a sufficient translation speed (around 110 words/min). The user interface is nicely done with a special purpose desktop and the possibility of translating multiple files in batch mode. The system can be called from Word to send texts directly to the translation engine.

Power Translator Professional is a system by Globalink which does high speed translation (up to 500 words/min). Of course this speed is traded in for accuracy in translation. While in general its user interface is convenient with a great flexibility in window layout, there are shortcomings in the module for adding dictionary entries. Adding such an entry requires the lookup of numerical codes.

Personal Translator was developed by IBM and is distributed by Rheinbaben & Busch Verlag. It is built around a linguistic grammar theory (i.e. slot grammar) and strives towards high quality translation. This results in a very slow performance (less than 50 words/min). It is the only system in this price range that includes a translation memory module. Unfortunately, the translation memory and the translation engine are poorly integrated. Personal Translator can be interfaced to Word which allows for texts in the word-processor being sent to the translation system.

Telegraph is a new system by Globalink. Its translation quality has been improved considerably compared to the German Assistant and Power Translator products. From the German Assistant it has inherited the option for interactive translation. In addition, it includes the very attractive feature of adding translation rules, thus enabling a better customization.

Transcend was developed by Intergraph and is distributed by Trados as part of their Translator's Workbench. The Workbench focusses on modules for translation memory and terminology administration. The Transcend system is only a side product. It translates with reasonable quality and good speed but it lacks a convincing user interface.

All of these systems translate a wide variety of grammatical phenomena as long as these are not buried in too complex a sentence and the vocabulary is all known. They manage for example word order very well, recognize most questions and commands and, in the case of German, combine separated prefixes with verbs in order to find the correct translation. More exotic phenomena such as postpositions (which are rare in German) or the fronting of special constituents can only be resolved by some of the systems. Only one of the systems (IBM's Personal Translator) attempts anaphora resolution across sentence boundaries. But the use of this feature is discouraged because it will slow down translation speed too much.

Summing up: translation systems which are available in the low price range are not perfect but they can be used profitably if their constraints are recognized and if they are employed for the appropriate tasks.

3 Criteria for the Employment of Machine Translation Systems

Machine translation systems can be profitably employed in a commercial setting if most of the following criteria are met. Our list is based on results reported in [ABH+94], [New92], [HS92] and [Nir93].

3.1 Linguistic Criteria

MT can be used if the source texts consist of repetitive and definable vocabulary. That means that the system's vocabulary can be tailored to the sublanguage vocabulary of the texts. This may take some time but it will substantially improve the translation results.

The source texts should originate from a clearly defined subject area. Many translation systems allow for the selection of a subject area before the translation starts. In this way the translation of homonyms can be influenced. If, for instance, the user selects the subject area 'banking and finance' and the system is to translate the German word 'Bank' it will correctly choose 'bank' rather than 'bench'.

Furthermore, the source texts should be stilistically simple and avoid ambiguous words and constructs. The handbooks accompanying machine translation systems give valuable advice on how to optimize the source text. They recommend short sentences, sticking to canonical word order (subject, verb, objects), the splitting up of complex attributes into multiple sentences, and the avoidance of nested structures.

The source texts may contain only few grammatical and spelling errors. The former are sometimes hard to detect and become evident only through translation problems. The latter can be checked and corrected with the help of a spelling module. It is obvious that a MT system has to rely on exact spelling much more than a human translator who can compensate this kind of problems by her text understanding.

3.2 Organisational Criteria

The translator's job is fundamentally changed by the employment of a machine translation system. Preediting and postediting become more central than actual translating. Preediting should be supported by special tools that check for problem areas in the source text. Problems can arise for example with abbreviations - a formal aspect which is difficult to handle for MT systems - or lexical and grammatical ambiguities. Some companies have even moved towards using controlled natural language, restricting the source text to a limited vocabulary with well-defined meanings and to sentences with a restriction on complexity and length.

In postediting it is important that the translator does not let the translation system influence her linguistic competence too much. She should not let the system lower her standards in translation quality. It is therefore important that she is able to influence the translation process in order to avoid repetitive correction tasks. Furthermore, familiarity with the system helps the translator to recognize error patterns and allows for faster postediting.

Much time and care has to be spent on updating the terminology database and the system's dictionaries. The translator has to become an expert in tuning the translation system. Therefore, she must understand how the system works. This entails that the translating staff must be thoroughly trained on the system. Ideally there should be computer experts, linguistic experts and translation experts working together.

3.3 Technical Criteria

First, the hardware requirements of the machine translation system must be met. Fortunately, powerful state-of-the-art PCs provide enough power (i.e. RAM and clock speed) for realistic employment of machine translation. Moreover, the hard disk requirements may not be underestimated. A system can use more than 100 Mbyte of disk space.

Next, texts must be available in machine-readable form. This will be ever more common in the future, however, it is not yet a matter of course. Paper copies must be scanned and converted to text with an OCR program (optical character recognition) which may introduce errors that have to be eliminated before translation starts.

In addition, the source text must be in a format that is compatible with the machine translation system. (Otherwise some filters have to ensure compatibility.) Fortunately, the PC translation systems support a variety of input formats, among them Word, WordPerfect and AmiPro. This is also important for the preservation of formatting information. The translation system should automatically transfer the layout format from the source text to the target text.

A multi-user environment increases the profitability of a machine translation system. It allows translators to share dictionaries, terminology databases and translation memories. Obviously, this requires careful access and version control of all parts of dynamically modifyable data.

Last but not least, support by the system distributor must be guaranteed. The software company should have a certain size, stability and continuity. Thus, the customer can be quite certain that the system it purchased will be maintained and extended in future years.

3.4 Psychological Criteria

Introducing machine translation will have a severe impact on the translator's work process. Therefore the psychological aspects may not be underestimated. Machine translation can only be successfully applied if the staff concerned supports this move. Towards this goal, extensive information must be provided in advance to dissipate the translators' possible mental reservations. The machine translation system must be seen as a tool, not a competition or menace. In the beginning the translator must go through comprehensive training so that she understands the translation system. She must then be given control over the translation process in order to cut out repetitive errors. Unfortunately, only few systems currently support such intervention.

3.5 Economic Criteria

Introducing machine translation will involve high initial costs caused not only by purchasing the translation system but much more by training the staff and by adapting the system to the user's specific needs (i.e. vocabulary, phrases and grammar). To justify such an investment there must be a continuing demand for and volume of translation. Entering into machine translation requires a long-term commitment. Users report that it may take up to a year before a real gain in productivity becomes visible.

4 Conclusion

It is clear that machine translation systems can only be employed with certain kinds of texts (e.g. technical documentation) and that their merits increase if they are used for large volumes of text with a definable vocabulary and subject area and with sentence structures of low complexity.

Current machine translation systems deliver raw translations with considerable speed. Such a translation may be used for a quick content check by somebody not familiar with the source language. In this way it can be determined whether the text is worthy a more thorough treatment. The raw translation may also serve as the basis for postediting by a human translator. It is important that special tools for postediting (e.g. parallel scrolling of source and target text) are integrated in the translator's workbench.

Translation systems should supplement translation memory systems. There is no use in machine-translating slightly changed versions of a text time and again, having to repeatedly correct some of the same mistakes. The translation system should be used for the first translation, which is then postedited by a human translator and fed into translation memory. Subsequent translations can then be based on the memory component. If a sentence cannot be found in the memory it can be processed by the translation system and marked accordingly. Translation system and translation memory should thus work closely together. If the translation memory contains a sentence that is very similar to the actual sentence but not quite the same, the translation system should be able to fill in the difference. However, no current system provides this functionality, as far as we know.

Future machine translation systems will have to improve on translation quality. This requires that they strive to model sentence meaning in more detail. Therefore, we anticipate that more specific semantic features will become more important.

If one attempts to achieve fully automated machine translation in the short run, one is bound to restrict the source language. In a Siemens project, linguists have defined what has been called 'controlled German'[1]. They have developed rules for conditional clauses (e.g. stipulating that they have to be introduced by a certain conjunction), coordinate structures (e.g. the first conjunct must not be elliptical) and many more.

We have argued that machine translation can be profitable if the source text is appropriately preedited (and controlled language is preediting in its most extreme form). This kind of preparation becomes even more profitable when the text is automatically translated into several languages. We can assume that a machine translation of a German text needs the same preediting when being translated into English, French, Spanish or any other language.

Finally, machine translation will profit from content oriented markup schemes like SGML (Standard Generalized Markup Language). It is obvious that translation quality will improve if the system can distinguish between a heading line, an enumeration, running text and so on. Therefore, the combination of word processing, document management and machine translation will become more fruitful in the future.

References

[ABH+94] D. Arnold, L. Balkan, R. Lee Humphreys, S. Meijer, and L. Sadler. *Machine Translation. An introductory guide.* NCC Blackwell, Oxford, 1994.

[HS92] W.J. Hutchins and H.L. Somers. *An introduction to machine translation.* Academic Press, 1992.

[Hut86] W.J. Hutchins. *Machine translation: past, present, future.* Ellis Horwood Limited, Chichester, 1986.

[KAT95] M. Kugler, K. Ahmad, and G. Thurmair, editors. *Translator's Workbench. Tools and Terminology for Translation and Text Processing.* Springer Verlag, Berlin, 1995.

[New92] John Newton, editor. *Computers in translation.* Routledge, London, 1992.

[Nir93] Sergei Nirenburg, editor. *Progress in Machine Translation.* IOS Press, Amsterdam, 1993.

[1] We are not aware of any accessible publication of this project. We base our observations on a brochure called 'Siemens Dokumentations-Deutsch. Syntaxregeln für Technische Dokumentationen' given to us by S. Schachtl in fall 1995.