Machine Translation for Home and Business Users

Hubert Lehmann
IBM Deutschland Informationssysteme GmbH
Wissenschaftliches Zentrum Heidelberg
hlehmann@vnet.ibm.com

Abstract: Based on the long-standing experience with the project *Logic-Programming based Machine Translation (LMT),* IBM and the publisher von Rheinbaben & Busch have developed *Personal Translator®,* a machine translation system which is geared toward the needs of home and business users. From the point of view of technology, the challenge has been to make most of the functionality available to general users which was originally developed in a research environment and with professional users in mind. It will be explained how the market of MT for home and business users was perceived and what features were deemed necessary for such a system. The technology of LMT which in its core parts originates from M. McCord (IBM Research) will be outlined. It will then be shown how general users can be enabled to tailor a system such as *Personal Translator* to their needs, in particular how they can modify the vocabulary. A discussion of translation quality and its parameters will follow. Finally, an outlook will be given of what can be expected in future MT systems.

Introduction

When research on machine translation started in the late forties, the linguistic theory which would have been needed as a foundation for such an ambitious undertaking simply did not exist. Since then progress in linguistics and in Natural Language Processing has been immense, even though we still only partially understand the problem. When Michael McCord of IBM Research started the project Logic-programming based Machine Translation (LMT) in 1985, he was able to build on the latest insights from linguistics, logic, and logic programming. As a result we now have a machine translation system which exhibits the most modern technology commercially available today.

Research on LMT soon became an international effort with close and enthusiastic cooperation of IBM researchers in the USA, Denmark, Egypt, France, Germany, Israel, Italy, Japan, and Spain. Accordingly, a number of prototype MT systems were developed and tested with the translation of computer manuals and other types of texts. It could be shown that LMT was able to meet the need for machine translation, and a lot was learned about the requirements of different types of users and different types of translation tasks.

The development of a commercially viable product based on the existing prototype software has been mainly the responsibility of IBM Germany's Heidelberg Science Center, and it was done in cooperation with the publisher von Rheinbaben & Busch, who has been closely involved in the product design and now is responsible for the marketing of *Personal Translator.*

In the sections to follow the topics will be: the market for machine translation, requirements for a machine translation system, the technology of *Personal Translator,* its user interface, a discussion of translation quality. In closing, an outlook will be given of how the development of MT systems may continue, and what can be expected for the years to come.

The market for machine translation

The need for translation is driven by commercial and cultural exchange between people who speak different languages. Today this need far exceeds the capacities for conventional translation, and it is rapidly growing as commercial and cultural relations become stronger and stronger. To many people, machine translation looks like the only means to cope with this ever growing need for translation. Whether such a hope is realistic, crucially depends on the translation quality which MT systems are able to produce.

Current MT systems are still far from producing *fully automatic high-quality translations.* Although some of them have been in commercial use for a number of years now, they have not been able yet to significantly tap the market potential. From a marketing point of view, the first question is: Who can benefit from what today's MT systems do produce and under what circumstances? The second question, which is almost as important, is: How should the machine translation function best be embedded into the user's working environment?

It has become clear in recent years that there is a great diversity both in types of potential users of MT and in types and respective volumes of texts to be translated. At the one extreme there is a huge number of home users with translation needs of maybe a few pages per year and on the other extreme there are relatively few large international companies and organizations which have translation needs of hundreds of thousands of pages per year. It is hardly conceivable that a single product could be flexible enough to satisfy the needs of all of these potential users.

Types of users

Users can be classified according to the languages they speak, their profession, and the types of organizations they work in. The relative importance of languages depends on the respective numbers of speakers and the intensity of commercial and cultural relations. Thus it is no accident that language pairs such as German-English, English-French, or English-Japanese are most popular in current MT systems. It could be argued, however, that the availability of less common language pairs on the one hand could serve as a lever for stimulating trade relations, and on the other hand could help to preserve the vitality of "smaller" languages.

A classification of users according to their professions yields something like the following list:

- home users, e.g. pupils, students, learners of foreign languages, foreigners,

technology freaks,
- secretaries, clerks,
- executives,
- professionals: technicians, engineers, scientists, researchers, lawyers, architects, consultants, auditors, teachers, etc.,
- professional translators.

Experience has shown that only a part of translations are carried out by professional translators. Due to time and cost considerations, many translations are done by professionals of various kinds who would profit considerably from MT and translation tools. It is our assumption that this will actually be the key market for promoting MT. Professional translators are expected to follow as soon as they realize the potential for increased productivity and experience pressure from their customers.

Types of texts and translations

It will neither be possible nor necessary here to give an exhaustive list of types of texts and translations which are relevant for machine translation. I will use a few typical examples:

- *Business correspondence* is the most widespread type of text, of interest to all types of users. It is characterized in part by stereotyped language (which often can be captured by a translation memory), by industry-specific terminology, and by very mixed linguistic and stylistic quality. It makes an important difference whether incoming mail is to be translated for informatory purposes or whether outgoing mail is to be translated. The latter usually will need to be postedited by someone who knows the target language.

- *Technical documentation* makes up most of the potential translation volume. It requires careful preparation with respect to terminology work and can profit enormously from linguistic or stylistic preprocessing (which may also improve the readability of the original for human readers).

- *Scientific publications* in foreign languages can become more quickly accessible, if MT is used to enable "information scanning". This presupposes however the availability of suitable terminological dictionaries, since otherwise, even raw translations may be too misleading or incomprehensible.

Requirements for a machine translation system

Ideally, machine translation is a batch process which is applied to a given text and which produces a perfect translated text which then only needs to be printed out. There are two observations to be made, however:

1. Texts come in many different forms and formats. They are in general not simply well-behaved sequences of grammatically and orthographically perfect sentences.

2. MT systems are not perfect:
    a. Dictionaries are incomplete.
    b. Different subject areas require different translations.
    c. Different styles may require different translations.
    d. Some sentences cannot be analyzed correctly, and hence they cannot be translated correctly.
    e. Some ambiguities cannot be resolved correctly.
    f. Intersentential dependencies cannot be detected or handled correctly.

A practical MT system must take these factors into account. The system must be properly embedded into the user's working environment, which means first of all that there must be a workable interface to text processing systems, preferably to the one the user likes best. For small texts, it may not be all that important whether and how layout information is treated, but for longer texts with complicated layouts this may decide whether the MT system can be used at all.

Dictionaries will always be incomplete. This means that users must be enabled to edit and improve dictionaries, and to specify all (or most) of the information which is required by the MT system to produce correct translations.

Users must be able to specify the subject area(s) which a text is about. However, the subject area of the text will in general not determine the translation of every instance of a term. Even when the subject area would be specified for each sentence, there could still be different translations required as in *List the orders in alphabetical order.*

Stylistic differences may be lexical, i.e. largely a matter of proper specification in the dictionary or they may involve forms of address (e.g. personal letters vs. official correspondence) or other phenomena such as the rendering of imperatives in manuals.

Where the MT system fails to produce the proper analysis due to internal deficiencies or due to improper choices for ambiguity resolution, the user must be able to postedit the translation. The necessary editing tools should preferably be integrated with the usual text processing system.

The technology of *Personal Translator*

Architecture
Personal Translator has been designed to run on PCs under Windows 3.1 or later. It may also be run as a Windows application under OS/2 3.0 (Warp). It requires at least 8 MB main memory and needs about 15 MB space on the hard disk. The system translates from German to English and vice versa.

Personal Translator consists of the following components:

1. MT engine,
2. dictionary component,

3. translation memory,
4. user interface,
5. interface to MS Word for Windows.

The MT engine is in essence the LMT system (McCord, 1989). The LMT system is built on a *Modular Transfer* approach, it consists of a language-independent shell (e.g. parser), monolingual components (e.g. analysis morphology and grammars), and bilingual components (structural transfer rules).

Source analysis
The analysis of the source language is based on *Slot Grammar* (McCord, 1980, 1991), which is both theoretically well-founded and suitable for efficient implementation. Slot Grammar views sentences and their constituents as *heads* which have *slots* for *modifiers.* The analysis of a sentence thus describes, which modifiers *fill* which slots of the heads present. Grammar rules specify 1. under what conditions a given modifier may fill a slot, 2. what are allowed orders of modifiers, 3. what slots may be extraposed in complex sentences, 4. what modifiers are accepted as adjuncts.

Before grammatical analysis takes place,

1. the source text is separated into sentences (respecting abbreviations, dates, numbers, etc.),
2. sentences are tokenized,
3. tokens are morphologically analyzed and looked up in system-supplied and user dictionaries,
4. multi-word expressions are recognized,
5. unlikely readings of words and expressions are eliminated to reduce the search space.

Grammatical analysis is performed by a chart parser. A parse evaluation algorithm is applied whenever a new node is constructed, to eliminate improbable structures as soon as possible. The parse trees which are constructed encode both a linear structure of constituents and a dependency structure which is independent of word order.

After parsing has been completed, references of anaphoric pronouns are resolved using an algorithm developed by Lappin and Leass (1994). Thus the correct translation of pronouns can be achieved.

Transfer and target language generation
Transfer takes the syntactic representations which are the result of parsing a source language sentence, and proceeds in two steps: 1. lexical transfer (already taking contextual restrictions into account) and 2. structural transfer which is performed using a set of transformation rules.

Lexical transfer substitutes words in terminal nodes of the source language parse tree by expressions of the target language (single words or multi-word expressions). Thereby translation conditions are respected which are either slot

conditions, path conditions, or global conditions. Slot conditions specify the presence or absence of certain fillers and arbitrary properties of such fillers (e.g. existence of a specific semantic type). Path conditions specify arbitrary relations between constituents and formulate any conditions on such constituents. Global conditions specify subject areas, general preferences, and can also be used for general regional or stylistic choices.

The target word or expression also receives grammatical properties which are computed from the grammatical properties of the corresponding source language constituent and change specifications. Change specifications may involve grammatical features, part of speech, slots and fillers, or arbitrary lexical transformations.

After lexical transfer, the syntax tree is restructured to reflect the correct syntax of the target language sentence. This is done by applying structural and lexical transformations in a predefined order. Transformations may move, delete, or copy nodes of the syntax tree, and they may contain arbitrary tests to restrict their application. A typical structural transformation is the one which handles the clause final position of verbs in German dependent clauses when translating from English. A typical lexical transformation is the one which turns English participial clauses into German daß-clauses.

When structural transfer has been completed, the correct morphological forms of words are generated, and finally the result is linearized and returned to the user interlace.

Dictionary component
Personal Translator uses three dictionaries: 1. high-frequent and highly complex words, which are kept in core memory, 2. the standard dictionary of approximately 50,000 stems per translation direction, 3. the user dictionary. The dictionaries are stored in direct access files approximately in the form described in McCord (1989). The dictionary entries contain specifications of slots and fillers, transfer conditions, and descriptions of structural change.

In user dictionaries, it is possible to define nouns, proper names, adjectives, and verbs, and to restrict translations to specific subject areas. The full complexity of the dictionary formalism of LMT is not made available, in order to make dictionary update as simple as possible for the user.

Translation memory
Source language sentences and their translations are stored in a database for further use. They can be retrieved using a similarity search algorithm. In *Personal Translator,* this is done from within MS Word for Windows. For each sentence of a text to be translated, the database is searched for 1. existing manual translations and 2. for translations produced by the MT engine. The user can use these translations and modify them as desired.

Formatting information is not stored in the translation memory, but is derived from the source language text. This has the advantage that formatting information need

not be considered during the automatic translation. However, in some cases of intrasentential formatting, the information may get lost.

Integration with text processing system
Personal Translator can be invoked directly from within MS Word for Windows, either by specifying that the whole document shall be translated or by specifying that a marked portion of text shall be translated. Before the translation is started, translation options such as subject areas to be used can be set.

User interface

The user interface of *Personal Translator has* two modes: 1. stand-alone and 2. integrated with MS Word for Windows. The interface is designed according to the principles of Common User Access (CUA) in order to make the interaction as intuitive as possible for anyone familiar with Windows.

There are two windows for holding the source and target texts. Files can be loaded into these windows, and the contents of the windows can be edited and saved in files using standard functions and dialogues. All of the source text or marked portions of it can be translated. During translation, a progress indicator tells which sentence is currently translated, and in which phase the translation is. An ongoing translation may be interrupted at any time, e.g. when an error is detected in the source text. Unknown words and translation errors are reported in a message window.

The following translation options can be specified:

subject areas,
impersonal imperative (translation of English imperatives as infinitive constructions),
translation of "you" by "du" instead of "Sie"
translation of German "Sie" at the beginning of sentences as "you" or "they",
resolution of pronoun references,
automatic decompounding of unknown German words,
interpretation of line end characters as terminating sentences,
input of "ss" instead of "ß" (primarily for Swiss users),
output of "ss" instead of "ß" (primarily for Swiss users),
time limit for translations.

Further options allow the specification of the files used to access the translation memory and the file holding the user dictionary.

The dialog for updating the user dictionary can be invoked in different ways, but most conveniently by marking a word to be added or changed and then pressing F11. Due to a careful choice of default assumptions, the amount of information which has to be specified by the user is kept to a minimum. In addition, the required information is prompted via examples such that users can directly apply their intuitive linguistic knowledge without worrying about linguistic terminology. Online help explains what the information prompted for means and how it is used.

Translation quality

Translation quality usually is measured in terms of percentage of correctly translated sentences with respect to the total number of sentences in a given text. Unfortunately this measure does not mean very much, as it is possible to construct texts such that for a given MT system the translation quality will be near 100% or near 0%. For a proper assessment, we need to consider the translation quality achieved when all words are known to the system, we need to exclude or correct sentences which contain orthographic or grammatical mistakes. For texts picked at random from the areas of business correspondence and computer manuals, *Personal Translator* can be expected achieve between 70 and 85% correct translations under these circumstances.

The remaining types of errors either are intrinsic to the system or at least cannot be corrected by a user's tuning of the system. These errors inevitably will be

1. ambiguities which are not correctly resolved,
2. syntactic constructions which cannot be analyzed,
3. too great overall complexity of sentences,
4. generation of ungrammatical target structures,
5. ambiguities which are generated, but which are not present in the original,
6. translations which are correct in isolation but not in context.

For practical purposes, we need to consider the effort which it takes to define all unknown words and expressions. For a short text this can be more time consuming than doing the translation by hand. Therefore we need a second measure, namely the ratio of unknown words to total number of words for a given text. This type of data is not yet available for Personal Translator, however.

Outlook

Practical MT systems have been around for a number of years now, with mixed success. With *Personal Translator we* hope to learn how machine translation can become useful to a large variety of users, and certainly many areas for improvement will become apparent. We know already that the current technology has not been pushed to its limits yet. Dictionaries can be considerably improved given sufficient resources. The ability of computers to handle huge masses of data has not really been exploited with current dictionaries. The linguistic analysis of really large corpora has just begun, and it may revolutionize lexicographic work and the quality of dictionaries we can make available to MT systems. Thus we can expect that in the foreseeable future we will have MT systems which have a better grasp of vocabulary and in particular specialized terminology than most human translators.

We can also expect advances in semantic and pragmatic processing of texts, which will also result in a better translation quality. However, it is much harder to predict exactly how and when these advances will happen. And it is clear that with

respect to semantics and pragmatics, human translators will remain superior to machines for a long time to come.

Acknowledgements

Of the many people who have contributed to the LMT project and the development of *Personal Translator,* I would like to thank especially Michael McCord, Brigitte Barnett, Arendse Bernth, Brigitte Bläser, Peter Blok, Lisa Breidt, Kurt Eberle, Herb Leass, Susan McCormick, Bernhard Keck, Annett Laube, Sonja Müller-Landmann, Mary Neff, Nikolaus Ott, Ulrike Rackow, Ulrike Schwall, Angelika Storrer, Amelie Stephan, Claus Stumpp, Christiane Sturm, Andrea Zielinski, and Magdalena Zoeppritz for their technical contributions and Peter Bosch, Reinhard Busch, Anne-Marie Derouault, Wolfgang Glatthaar, Otthein Herzog, and Jaap Hoepelman for their managerial support.

References

Lappin, S., H. Leass (1994): "An Algorithm for Pronominal Anaphora Resolution", *Computational Linguistics,* 20, 535-562.

McCord, M. (1980): "Slot Grammars", *Computational Linguistics,* 6, 31-43.

McCord, M. (1989): "Design of LMT: A Prolog-Based Machine Translation System", *Computational Linguistics,* 15, 33-52.

McCord, M. (1991): "The Slot Grammar System", in: J. Wedekind and C. Rohrer, (eds.): *Unification in Grammar,* MIT Press. (To appear).