Linguistic-Technical Aspects of Machine Translation

Harald H. Zimmermann
Universität des Saarlandes
5.5 Informationswissenschaft
D-6600 Saarbrücken 11
FRG

(erschienen in: AGARD Conference Proceedings No. 430, Washington D.C., 1987)

Summary

To allow to compare computer aided translation (CAT) and machine translation (MT) systems, essential criteria and typical exponents of the various concepts are presented. Among the most important criteria of differentiation are the following:

- Translation of complete texts (including titles, abstracts) vs. lexical aid for translation
- All-round translation vs. translation of specialized fields and texts
- Bilingual translation vs. multilingual translation
- User-oriented vs. administrator-oriented expansion of dictionaries and systems
- Portability.

Included essential criteria for evaluation are:

- Quality of the raw (also called informative) translation; advantages / disadvantages as opposed to human translation
- Embedding in a user environment (data bases, indexing and retrieval, word processing, electronic publishing)
- User friendliness
- Costs of development and installation
- Running costs and effects of realization.

List of symbols

CAT	Computer aided Translation
EC	European Community
EUROTRA	European Translation System, research project of the European Community
LOGOS	Commercially available Machine Translation System of LOGOS Corporation
MINITEL	Videotex system access tool to the French TELETEL
MT	Machine Translation
PC	Personal Computer
SUSY/STS	Saarbrücken Translation System / Service
SYSTRAN	Commercially available Machine Translation System of World Translation Center, U.S.A., and Gachot S.A., France
Wang OIS	Office Information System of Wang Computers
WordPerfect	Commercially available editor

Introduction

Considering the technical development in information industry, one has to realize that the computer will be integrated in nearly every professional human activity. In the field of word or text processing, man-machine interaction, on PC level or on the base of a work station on terminal instead of a typewriter, will be the standard in the near future, that means from the end of the eighties.

On the other side, the problem of automatic translation of natural text in general - even not to speak about speech understanding and speech translation - will not be solved in the sense of FAHQT (fully automatic high quality machine translation), due to the complexity of natural languages. So there are limits (more or less in the scientific approach) and possibilities of integrating technical tools in the process of (human) translation or even text understanding by machines.

The aim of this contribution is to explore the possibilities of usage of the computer in the field of machine and machine aided translation. Even if it might be of some attraction to treat also principal problems of translatability of texts (and understanding), we will concentrate on technical translation, that means the translation of technical or common texts.

Systematic aspects

Even if it might be of some interest how machine translation is realized, the linguistic aspects (esp. the models of grammar and the possible or used strategies) play a subordinate part. So one can assume that there is a kind of "black box", where a text or words in one natural language are put in and with or without human interaction one or alternative translations come out of the system. The translation might be good, useful or bad under the aspects of the user.

Concentrating on the usage and usefulness of machine translation or computer aided translation, one has - in principle - to distinguish between two main user groups:

(1) the so-called end user, e.g. an expert trying to get the information of an article written in a natural language more or less unknown to him; a person writing a letter to a friend in a foreign language ... and

(2) a professional agent, esp. a human translator who tries to use the machine as a tool in the process of fulfilling his job.

Under these aspects, one can let aside projects and basic research on the principles of machine translation / language understanding and orientate on practical tools resp. systems.

As a starting point, it is important to distinguish between two main strategies: machine translation (MT) and computer aided translation (CAT).

- A system will be called a machine translation system (MT) only if a translation process - starting from the machine readable source text - is fulfilled without any human interaction to reach a target text quality at least "good enough" for information purposes.

- A system will be called a computer aided translation system (CAT) if a human interaction is needed or foreseen to reach the aim of a "good" translation of a source text (machine readable or not).

It is quite clear that an MT system can be used, in addition, as a component of CAT: a text can be adjusted before the MT starts to get better machine translations ("pre-editing"), and/or a machine translated text can be post-edited by human translators to get higher quality.

There are a lot of systems on MT and CAT on the market claiming to be the right practical tool, and the choice is hard to be made without having precise criteria for the decision.

Because the amount of data in computerized dictionaries - on the long line of development of MT and CAT systems - is the decisive component, the update of the (electronic) dictionary plays a substantial part in both alternatives:

- One concept may be - especially if the system still has a large capacity of dictionary entries - that the user, e.g. the translator should not take part in such a process (see, for example the SYSTRAN concept), so that there are specialists needed on the system administration side to do the job of improving the dictionary data base;

- the other way is to leave it to the professional user to complete the system's dictionary or to add a special vocabulary (see, as an example, the LOGOS concept).

There are other system aspects which play an important part in the decision process: the availability of language pairs, the possibility of using or handling special text types (e.g. minutes, letters, ...).

Criteria of evaluation

Quality

Even if one normally can (or has to) handle the system as a "black box", there are differences in the quality of the "pure" machine (raw) translation results (output). It is not very easy to give a precise measure, but there are some important criteria to be noted (see, for details, esp. the concept of Van SLYPE (1982) and the descriptions of the two SYSTRAN evaluations, Van SLYPE (1979)).

The main criteria under these aspects are:

(1) reliability and fidelity, that means: to what degree (bad, good enough, good) the content / meaning of the original text is conserved.

(2) understandability and intelligibility, that means: to what degree the end user is able to read and understand the translated text.

This looks quite simple, but the problems occur in the detail: So, on the syntactic and stylistic level, a system normally will produce translations which are worse than human translations, whereas on the lexical level, esp. in the identification of the right technical term(s), a system's translation may even be more precise and consistent than a human translation.

Application environment

MT or CAT has to be seen in the (technical) environment of application possibilities. So, normally a decision is not only quality oriented, but also based on the possibilities of integration into a complex text or word processing system. Under these aspects, the following components have to be considered:

Integration in (bibliographical or textual) data bases

Meanwhile, data bases are - technically spoken - world-wide accessible via packet switching networks and even satellite communication. So, overcoming the language barriers, e.g. between English and Japanese, but especially in the European multilingual market, becomes an important desire. Experimental efforts to integrate MT systems in such an information process are made in Japan: as one example the INSPEC data base which is originally English can be accessed with Japanese key words. The key words and later the (English) title are automatically translated during the dialogue into Japanese (Nagao et al. 1982). In a similar way, an MT system for German to English is used via a batch process and with post-editing to translate the titles of German data bases (Zimmermann et al. 1987).

It is quite clear that translation of titles and abstracts could also be done by human translators. But there are some arguments for MT and CAT: the text to be translated is machine readable, so there is an ideal base for using a computer, the fields or areas on which the title / text is oriented are normally "physically" marked, so that the classification or even thesaurus functions can be used especially for the lexical transfer (disambiguation), the technical vocabulary needs to be very precise, so that the computer helps in being consistent.

Automatic Indexing

Indexing of (full) text might be a good side-effect of using MT and CAT. For lexical transfer, one needs to derive word forms to basic forms; compounds and complex words have to be identified as such, word class information, even relations between terms are used for disambiguation purposes. So it can be considered to provide such output or intermediate results of MT systems for the purpose of document archiving and information retrieval.

Text and word processing

There is no doubt that text processing plays an important part in every translation environment. Even free-lance (human) translators more and more will use a word processing system (on PC), and there is a small step to integrate, in any way, (private) glossaries or word lists accessible via so-called "windows" on the screen - instead of using card-index boxes. It is quite clear that other functions, e.g. spelling, grammar and style checkers, will increasingly be integrated in such a process.

As a result, the "source text" is machine readable. But MT and CAT system have to be adapted to the (different) word processing systems (for example the Wang OIS is combined with LOGOS and SYSTRAN, WordPerfect is combined with SYSTRAN). If such tools are available, post-editing of MT results can be supported by the special editor.

One problem in this environment is the combination of those tools with local facilities of MT (see, e.g., LOGOS) or the connection with a Translation Service Center (see, e.g., the concept of the SYSTRAN application in the European Commission or even the use of the MINITEL-System in France to get machine translations (by SYSTRAN)).

There is no doubt that the usage of MT and CAT will make progress mainly in combination with text processing and online access. The question is, for the moment, if the existing tools are powerful enough (in quality) that they will be accepted by the user. Especially the results of the MT experiment done by Gachot S.A. on MINITEL and on PC-Access (with SYSTRAN) will be an important contribution in this direction.

Electronic Publishing

More and more, texts to be translated (esp. technical text like repair instructions, manuals) are fully prepared - including figures, drawings, tables, pictures - via electronic publishing, nowadays also known under the variant of desk top publishing. Companies giving commissions to translators - inside or outside the firm - don't want to rearrange (or compose) the complete product of the translation for even a lot of target languages.

So a great effort has to be undertaken (may be, on both sides: the producer of electronic publishing systems and the producer of MT and / or CAT software tools) to integrate translation helps without deleting or violating the document structure. Of course, there are problems in line, paragraph or page adjustment (due to the different lengths of translated text), and also the rearrangement of phrases / words due to the different word order raises problems on the correct integration of typographical markers (bold face, underlining etc.) in the target language. If these "technical" components of translation processes will not to be left to the human translator / posteditor, a higher standardized level text description must be integrated in the electronic publishing facilities. This will be a great challenge to the existing and coming MT and CAT systems.

User friendliness

On behalf of the professional translator's work station, user friendliness plays an important part on every "level" of MT and CAT. Nobody should want that in the use of MT or CAT the human actor has to play the part of a "slave", e.g., correcting, day by day, only the "trivial" mistakes of the system's output. There is, at the moment, such a danger, because the existing systems are not very flexible and adaptive.

Future developments in MT and CAT must therefore concentrate on activities which give the user more and direct feedback possibilities. The "private file concept" which sometimes is used in data base systems, where a user can select and create a "personal part" in a data base, could be an example: at least on the dictionary level the user should get functions to realize - on the basis of the existing data - his "own" dictionaries (physically or logically).

By the way, both sides - the provider of the data base system and the user - can make profit of such a concept: the system's specialized vocabulary will be ameliorated and the user will have a great (but also responsible) influence in the choice of the translations.

What is true for the lexical part should also be applicable to structural components (e.g. to influence the length of sentences, some stylistic components etc). Existing systems should get more flexibility and coming systems should consider such components from the beginning.

Cost and benefit

As everywhere in business, the decision of using a tool like MT or CAT is made on a cost and benefit calculation. In this case, it is not only a pure monetary problem, because "time is money" and getting translation without delay might be paid on a higher price. But at the end, the decision is made on an economic basis, having in mind the social or human effects.

Because there is no substantial data available, at least to the author, on the cost of development of MT and CAT systems, one has to concentrate on the cost and benefit on the user's side.

It seems, at the moment, that the amount of translations (measured in pages / day) which are produced by a human translator via interaction and / or post-editing can be substantially increased. Assuming that, for example, the maintenance and technical application of a system like SYSTRAN will cost $300.000/year, and that 300.000 pages can be technically translated by the system, the cost of the pure translation - not including the updating of the dictionaries and the preparing or post-editing of the MT results - nearly can be neglected ($1 / page). The cost of the complete process (translation with man-machine interacting) differs depending on the quality one wants to get. For a "good-enough" translation - which means informative translation, say, for example, for working papers, a so-called rapid post-editing can be done, so that a human translator produces about 20 pages / day (instead of 4-6 pages without any MT).

To reach high quality comparable to professional human translation, one has to consider more time for revising or post-editing. But it seems, that the break-even point is reached in the sense that if the vocabulary of an MT system is adapted to the user's field, the cost of translation is less than the cost of pure human translation, even considering that the use of word or text processing systems in the translation process saves about 20 % of time.

Basic linguistic and strategic problems

The morphological component (that means problems of inflection, derivation, and decomposition) in MT has been successfully solved, at least for practical purposes in the application environment, even if the correct translation of identified derived and decomposed words not always is reached via automated rules. This is not the case for solutions on the syntactic or semantic level. Even if we assume that a problem, e.g. the disambiguation of syntactic homographs (like RAINS in IT RAINS or THE RAINS) can be solved via a strict and fully formalized parsing system, the complexity of natural language structure leads to an explosion of computer time if one tries to integrate or handle every possible (partial) structure or occurrence. So, in reality, commercially available systems try to shorten the process of identification (or disambiguation) via special deterministic rules or probabilities. As a result, they run 10 000 or 1 000 times faster than a fully linguistic-oriented system, but their results may not reach the same quality.

Today, computer time plays not any more the same role as some years ago, but in machine translation computer time up to now is not fully neglectable. The same is true for the solution of problems of homonymy, that means in the semantic field. On the one side, there are limits in the handling of text structure vs. sentence structure. In most cases, the knowledge base of an MT system is the sentence environment, that means that information or solutions of previous sentences are lost and that nearly no data is known on the text level. This leads to many problems, especially in the field of pronominal reference, but also in article insertion and homonym disambiguation. The way existing systems handle the general problem of semantic ambiguities is by introducing semantic codes (which - on a general level - also plays some role in the disambiguation of syntactic structures), especially "field" or discipline markers, which are used to select the "right" word (or even word sequence) depending on field parameters given by the user. They also try to solve this problem by dictionary look-up to identify word sequences or even sayings (which normally have to be lexicalized because there exists no algorithmic solution, see the German ES REGNET BINDFÄDEN, which has to be translated by IT RAINS CATS AND DOGS.

Since Chomsky, the systematic structural-semantic access to language analysis, "understanding" and translation has made some progress. So, on the research level, there are several modern formalized grammar types and parsing systems available. Especially in Japan (see, e.g., the MU-System) and Europe (e.g. the efforts made by the European Community and their member states with the European Translation System, EUROTRA) research in MT is continuing. But it seems that one needs more than just computational linguistic development: linguists, computer specialists, information scientists and users have to cooperate in large-scale projects to reach the aim of practical usability.

Examples

To give some impressions of the state of the art of so-called productive (not to say commercial) systems, and also to show how the mentioned criteria can be applied, two systems, the MT systems SYSTRAN and SUSY/STS will be described.

SYSTRAN

SYSTRAN (commercial rights at Gachot S.A., France) - in its newest version 3.7 - has the following characteristics:

- Translation of full text. Even if the structure is not correct or words are misspelled or words are not found in the machine dictionary, a translation is produced.

- The speed of translation (depending on computer capacity) reaches up to 350.000 words per hour. So it is the fastest system available on the market.

- The system is language-pair oriented. Translations are available for the pairs English -> French, English -> Italian, French -> English, Russian -> English (USAF), English -> Japanese (SYSTRAN JAPAN), English -> Arabic; under development are, besides others, English -> German, French -> German, German ->English and German -> French. The quality depends, on the one side, on the availability of (discipline-oriented) dictionaries. A great effort has been made at the European Community to develop the SYSTRAN dictionaries. For the translation from English to French, the dictionary now contains more than 150,000 entries. The same quality is not reached, e.g., for translation from German to French, which is in the starting phase, even on the level of language analysis.

- SYSTRAN application needs technical specialists (and administration). So only companies which are able to realize a special staff (e.g. the EC or USAF) have the possibility to use a SYSTRAN version on their own computer (if the computer is a mainframe IBM or IBM-compatible). But there is an interesting alternative: to use the system via telecommunication networks, e.g. packet switching or - as a very "futuristic" example - via videotex. In France, a videotex application (using the French videotex version, called TELETEL, via a telephone combined with a screen, called MINITEL) is already available (and even used by pupils).

- Dictionary maintenance - up to now - for SYSTRAN dictionaries normally has to be done by system experts. The main problem is not the coding itself (which is very complex, but could and will be handled by user-friendly interfaces), but the consistency of the dictionary data base. Dictionaries contain area codes, but this component has to be developed to get a more flexible user and usage orientation.

- SYSTRAN is nearly not portable, that means: the basic code is an IBM-Assembler, even if the linguistic rules normally are written in a special macro language. The system itself needs - as mentioned above - a mainframe (IBM or SIEMENS or AMDAHL) computer or computers of similar size. A software revision (may be, on the base of UNIX) is planned.

- The quality of (raw) translation differs depending on the language pairs and dictionary satisfaction: If one concentrates on English -> French, the following percentages may apply: Morphological identification: about 100 %; syntactic structures: about 90 %; semantic disambiguation: between 80 and 90 %, depending on the integration of so-called limited-semantic rules.

- SYSTRAN can be combined with several word processing environments. One is the Wang-OIS (which is used at the EC), but one can also use PC (IBM compatible) with an editor like WordPerfect. There are applications where one uses an optical character reader to create a text file with WordPerfect, sending the data to the SYSTRAN service center (e.g. at Gachot S.A., Paris) and getting back the translated text via post line and a software tool which allows to post-edit the text having both versions - the original and the translation - on a split screen.

- The usage of the translation system itself is batch-oriented. During the process of translation, there is no possibility of intervention (despite the fact that the system administrator has the possibility to introduce translations for unknown words). On the other side, the user handles the system as a black box, he doesn't need any knowledge of the system.

- There is no information available about the development cost of the system. One can estimate that they are between $ 20,000,000 and $ 50,000,000. The EC application alone cost about $ 4,000,000 and $ 6,000,000. The cost running a version (without royalties) lie

- in my opinion - at $ 300,000 / year. So one can estimate that if one is able to translate about 300,000 pages a year (e.g. at the EC), that the system costs are less than $ 1 / page (word processing, post-editing and computer time not included). The translation fee on the French MINITEL (videotex) systems is about $ 0.10 per "window" (up to 10 lines), the automatic (raw) translation normally is available in this application within 45 to 60 seconds.

- This leads to the conclusion that SYSTRAN is end-user oriented, that means that its market is the "information society" where 100 % quality is not needed, but good-enough translation to understand a text written in a language not or nearly not known by the user. But there is no doubt that SYSTRAN has also a good chance to be used in the professional translation environment as a high speed tool and a possible alternative in computer-aided translation.

SUSY/STS

SUSY/STS, developed in the framework of a large research project in computational linguistics at the university of Saarbrücken (FRG), now is used as a production system in the field of translation of data bases. This service is performed by the Institute for Applied Information Research (IAI) at Saarbrücken. The main characteristics are:

- Translation of full texts (sentence level) is possible. In the application environment, one concentrates on title (and abstract) translation.

- Like SYSTRAN, SUSY/STS is an "all-round" and robust translation system: A text is translated even if there are unknown words included. In such a case, the original word is integrated in the text of the target language.

- To some extent, SUSY/STS is a "multilingual" MT system. Only the transfer component (between two languages) is bilingual, whereas analysis and synthesis are done independent from the target resp. source language. SUSY/STS applications are possible for the language directions German -> English, English -> German and Russian -> German, some effort has also been made for the development of French -> German and Esperanto -> German. The main application is made in German -> English, where a basic German analysis dictionary of about 150,000 entries, a German compound dictionary of about 140,000 entries, a German - > English dictionary of about 80,000 entries is available.

- Like SYSTRAN, SUSY/STS needs an administrative staff to handle the application. Therefore, one concentrates on a translation service concept. Translations are processed only at the computer center of the IAI, Saarbrücken. The clients send their data via magnetic tape to the IAI, they get back a magnetic tape including the translations. The main applications are, as mentioned, the translation of German titles coming from different data bases to English. Besides, the bilingual terminology in the different disciplines is updated, the dictionaries are augmented with the special terminology during the processing of the client's data. Clients, besides others, are: The Information Center for Building Construction, where a bilingual data base (ICONDA) is produced; the German Patent Office, where - at the moment - the German Catchword Index is translated into English; the German Institute of Standards, where the titles of German industrial standards are translated to English to be integrated in the corresponding data base; the Information Center for Social Science, where the titles of a bibliographical data base are translated. Up to now, more than 100,000 titles have been translated by this service.

- The System is available under UNIX (and in this sense, portable), the programming language is FORTRAN, with some basic parts written in "C". A version is also available on SIEMENS mainframe computer.

- The quality of the (raw) machine translation (German -> English) is satisfying: about 99 % identification of words in the source language on the morphological level; about 90 % correct identifications on the syntactic level. There are no statistics available about the correctness of semantic disambiguation, but development of tools for integrating discipline codes is under progress.

- The service normally integrates post-editing to reach a high quality of translation. As just mentioned, the clients send their data via magnetic tape to the service center. As a first step, a spelling check is done to identify and correct misspelled words. Then, the unknown words (mainly in the transfer dictionary of the MT system) are identified, the dictionary is then updated (by the human translator). In the next step, the automatic raw translation is done and the text is post-edited (on screen using a PC or terminal) by the human translator. After a controlling phase, the magnetic tape is sent back to the client. It is possible to integrate an indexing system, that means to use the analysis phase of SUSY/STS to produce lemmatized words (instead of word forms) and decompose or derive complex words to get relations between compounds and the single words from which they are derived.

- Like SYSTRAN, SUSY/STS merely is batch-oriented. But a special trained user would be able to integrate new words into the different dictionaries: the dictionary maintenance is dialogue-oriented. Post-edition can be done separately on PC or via terminal on the host system (i.e. the mainframe computer, which, at the service center, is a Nixdorf TARGON 35).

- The development cost of the system were about $ 5,000,000 (basic research at the university not included). The adaptation to the STS service environment cost about $ 200,000. The basic cost keeping the system available and running are about $ 150,000 / year, the service personal (not the translators) included. The service is cost-covering, if about 400,000 titles or sentences can be translated (fee / title: 2 DM = about $ 0.80).

- As a conclusion, one can say that SUSY/STS is a specialized CAT system. Its speed (5,000 running words / hour, CPU time, can be translated) is not comparable with SYSTRAN, and there will not be an augmentation of the language pairs. But it can be shown that, under consideration of the possibilities and by adaptation to special needs, MT systems will play an interesting role in the future.

Under the limits of this article, it is not possible to give similar descriptions of other existing (and available) systems. It also needs a good familiarity in the use and a good knowledge of the conceptual level. The aim was to give some concrete examples to give a first impression of the complexity of this theme. So it seems to be too simple to handle MT from the pure "linguistic" view (in the sense, that the quality of a MT is or is not comparable with the quality of human translation) or - on the other side - to handle MT as a pure technical software tool: Progress can only be made if one considers the limits and the possibilities of MT in concrete application environments.

5. Outlook

If one looks in a distant future, it seems quite clear that professional translation will normally be undertaken by using MT systems. It depends more or less on the availability of the "right" language pairs, the completeness of machine vocabularies, the availability of the "right" technical environment and, last not least, on the cost of the utilization of the system.

In-between (and besides), CAT systems on the level of vocabulary and terminology aid - integrated in or added to text processing systems, will play a substantial part, esp. as an aid during the creation of a text in a foreign language by the author or to understand a written machine readable text by the receiver arriving via telecommunication (Teletex). In the office environment, CAT, mainly with bilingual dictionaries on CD-ROM or hard disk will be used as a variant of style and grammar help.

Contrary to some assumptions (see, e.g., Hutchins 1986, pp. 331 - 334), we do not see any need for a special work station, but there will be special functions as translation help integrated in powerful writer's workstations, so that everybody will make profit of the development in computerized dictionaries and thesauri.

What will happen with the human professional translator in the environment of technical translations? One can see MT systems as special expert systems with the problem, that they will nearly never reach - generally spoken - the full quality of good and specialized human translators. But to develop, to augment and even to maintain such translation expert systems, one needs "language knowledge engineers" which are able to handle and to feed these systems. As far as the complexity of natural language and knowledge is concerned, there will be cooperation between the system and the translator. And there is a good chance, that if translation of texts will become cheaper and faster, more demand will arise, so that, in the end, human translators will still play a substantial part - under changed conditions.

Literature

Hutchins, W.J. (1986): Machine translation: past, present, future. Chichester.

Nagao, M. et al (1982): An English-Japanese machine translation system of the titles of scientific and engineering papers. In: COLING 1982, Amsterdam.

Van Slype, G. (1979): Deuxième Évaluation du système de traduction automatique SYSTRAN anglais-francais de la Commission des Communautés Européennes. Bruxelles.

Van Slype, G. (1982): Conception d'une mèthodologie générale d'évaluation de la traduction automatique. Multilingua 1, 221-237.

Zimmermann, H., Kroupa, E., Luckhardt, H.-D. (1987): STS - Das Saarbrücker Übersetzungssystem. Saarbrücken 1987.