WHITHER MT?

Bente Maegaard
Center for Sprogteknologi
Njalsgade 80, DK-2300 Copenhagen S, Denmark
E-mail: bente@cst.ku.dk
Phone: +45 35 32 90 90, Fax: +45 35 32 90 89

Abstract

MT started out as a 'technology push': more than 50 years ago, researchers had the bright idea of doing translation with the use of the newly developed computers. MT remained in the technology push area for many years. However, in the nineties we are seeing the 'market pull' beginning to play a role and there are good reasons to believe that this trend will continue.

MT is going where the market and the users wants it to go, and MT will be prospering in the future. MT will be available electronically over the network, and MT will be available in environments which also offer a variety of other tools for translation, as well as tools for other types of information management.

Also in research and in development of new technologies, MT will further develop, e.g. along the lines of knowledge-based MT, advanced integration of different analysis techniques (rule-based, statistics-based, etc.), integration with speech etc.

1. Introduction

MT is more than 50 years old [Boitet, 1995], and during the last 10 years it has reached a maturity which was probably not expected by so many people 15 years ago. Researchers and developers did believe in MT 15 years ago, and funding agencies believed in it too, otherwise the systems that we are currently using would not have existed.

Should we still turn to researchers when we want to know what will happen in the future? Yes and no. MT still contains so many unsolved problems that there is lots of research to be done in order to improve the results, and as can be seen below, many of the integrations with other technologies start out as research projects too. On the other hand, MT systems have reached a level where they are actually usable, and therefore some new developments may be largely introduced by groups other than researchers.

At recent conferences and workshops (e.g. the EAMT 1997 workshop in Copenhagen, proceedings in press) it has been remarkable too see that the audience now consists of highly experienced users of various tools. For the first time, users are able to express their views, not as an abstract wish, but as a fact because it builds on experience. This provides much better input for a discussion with producers. In fact, this provides at least *part of* a market pull. The most frequent problems for such users is the absence of integration of MT with other tools, and in particular of translation tools with document production tools.

Finally, availability over the Internet, integration with speech systems etc. gives a whole new range of possibilities which will also participate in directing the future of MT.

2. MT at present

2.1. Purpose of translation

For many years one concept of translation was prevailing: a fully proof-read high quality translation. And MT was supposed to provide an output which was at least close to that. At present, the situation has changed. First of all, MT is now expected to provide raw translation which is input for post-editing. This setting of the right expectations of course increases the possibilities of user satisfaction. But secondly, it has been realised that the old-fashioned concept of high quality translation for all texts was not adequate. Some texts are of a type that does not require in-depth post-editing; the European Commission has introduced the concept of 'improved' (améliorée) MT, i.e. slightly post-edited MT, for texts which are for information only.

Non-translators who use MT for information purposes, will not even post-edit the output, but use the raw translation. It is interesting to notice that the use of MT by non-translators is increasing and will further increase - non-translators will probably be responsible for the biggest increase in the use of MT, cf. the use of MT in the European Commission below.

Domain specific or general purpose?

The first translation systems were built around rather small corpora and therefore could be said to be domain specific. However, when they were further developed they were aimed at being general purpose, e.g. SYSTRAN. Later developments have shown the best results for restricted areas (TAUM-METEO, PaTrans), but there is clearly a trade-off between developing and maintaining a range of domain specific systems which give high savings and developing and maintaining one general purpose system. Probably the domain specific parts can be kept quite small and chosen by giving parameters in the system call.

2.2. An example: the European Commission translation service (SdT)

The European Commission in Brussels and Luxembourg has the world's largest translation department with some 1200 translators, 100 linguistic support staff, approximately 100 management staff and close to 500 secretaries and assistants [Theologitis 1997]. The SdT made a feasibility study in 1996 with a number of goals. The study investigated the users' degree of satisfaction, cost/benefit analysis, and the market outside the Commission. The purpose of the study was to give input for planning for the future, in particular is MT useful for the Commission?

The Commission has two types of users: the translators of the SdT who may use MT to provide the first step in the translation process, and the staff members of the other Directorates who may obtain a fast raw translation this way.

In the SdT, users of MT as well as non-users were asked. The main conclusion was that the user respondents in SdT found that it is useful to have access to MT. SdT produces quality translation as well as 'fast post-edited' translation (for use in cases where this is sufficient);

two thirds of the users used the MT output as input to quality translation. The users were very satisfied with the MT response time, and some of them would not have been able to finish their job in time without access to MT. I.e. altogether a very positive evaluation of the services available.

Non-users in SdT are probably not so interested in the question, so the reply rate was much less than for users (34% compared with 71%); the main reason for non-users not to use MT was missing information, and also the group of non-users contains translators who as a matter of principle do not want to use MT, as well as translators who do not find their language pair covered in the present situation.

The SdT staff use 30% of the MT used at the Commission whereas the other Directorates uses 70%. Consequently, the answers by staff members of the other Directorates are quite important. The main reasons to use MT are: first of all to translate documents which are less important, secondly to have documents translated fast (would have preferred to send the documents to SdT, but could not wait), and to use the raw translation for browsing.

Of course, some staff of the other Directorates do not use translation at all, other do the little translation they need by hand etc., but also in this case a major reason for not using MT was missing information about availability.

It was also part of this investigation to make a survey of the outside market to compare with the tools available at the Commission. The main conclusion was that those 6 language pairs which attract most of the positive feedback in the Commission will hardly meet any competition outside, and that for certain language pairs not available at the Commission, the market offers solutions. This is e.g. the case for certain Scandinavian and Nordic languages. (It should be noted that most of the translation in the Commission is between the 3 languages English, French and German).

The human effort reduction obtained by the SdT by using MT is calculated at an average of just over 35% of the time spent on a translation task. This is a very good result which should be seen in the light of the fact that SdT translated some 60,000 pages using MT in 1995 and probably 80,000 in 1996.

As can be seen from the above, the result of the feasibility study is extremely positive: users are satisfied and rentability has been proven. Consequently, the Commission will further pursue the use of MT, including the acquisition of more language pairs.

2.3. Another example: Lingtech

As another example, we will look at a much smaller enterprise, namely Lingtech, a Danish company specialising in translation of technical documents, in particular patent texts. Lingtech had a tailormade system made for them to be used in particular for patent translation English-Danish. Center for Sprogteknologi (Centre for Language Technology) developed the PaTrans system for Lingtech, on the basis of the EUROTRA prototype. In contrast to the systems used in the Commission, the PaTrans system is domain and text type specific. Consequently, Lingtech has been able to report savings of translator time of around 50% using MT [Maegaard and Hansen 1995]. At the same time, the turn-around time for a translation is decreased.

2.4. Conclusion on MT status

These two example are of course merely *examples;* there are plenty more. From this we can see that MT is at present sufficiently good to provide important savings both in terms of time and in terms of manpower.

However, as mentioned in the introduction, MT still has a major scope for improvement! This is true both technically and with respect to integration, exchange of data and workflow.

3. MT improvements proper

3.1. Interlingua and transfer, translation theory

Current rule-based MT uses either the transfer architecture or the interlingua architecture, mostly transfer.

Interlingua approach

ST **-> analysis ->** IL **-> synthesis ->** TT

Transfer approach

ST **-> analysis ->** IS(S) **-> transfer ->** IS(T) **-> synthesis ->**TT

where ST: source text, TT: target text, IL: interlingual representation, IS(S): interface structure representation for the source text, IS(T): interface structure representation for the target text. 'Text' normally stands for 'sentence'.

The interlingual approach requires a full analysis leading to an abstract representation which is independent of the source language, so that the synthesis of the target sentence can be made with no knowledge of what the source language was.

By contrast, the transfer approach uses the comparison between the two languages involved. The transfer phase exactly compares lexical units and syntactic structures and produces the target IS representation using the information, plus any additional semantic or other information available in dictionaries or knowledge bases. In the transfer approach, nothing is decided in a priori about the depth of analysis, i.e. the depth of analysis can depend on the closeness of the languages involved, the closer the languages, the shallower the analysis.

The transfer approach is often chosen simply because it is the simplest and most practical - this is the most important virtue in the development of production systems. However, researchers will pursue the beauty of the interlingual approach, if only because this type of research contributes to our understanding of what translation is. Discussion has been

ongoing if one can only translate when one has understood the text. This counts for humans as well as for machines. The attitude that understanding is necessary has many supporters.

This discussion: transfer-interlingua? understanding necessary? brings up a new aspect: type of translation. Traditionally, a distinction was made between 'literal' and 'free' translation, where it is obvious that certain text types require literal translation (e.g. EC texts) and others require free translation (e.g. literature). This informal distinction has been further refined in translation theory, see e.g. Reiss [Reiss 1976] for a discussion of the relationship between text type and translation type. In this theory, the communicative purpose of a text determines the type of translation it can take.

When experts in translation theory talk about translation as Textual (Re)production, they imply that in order to translate, the translator has to create the meaning of the source text in his/her brain and then express this meaning in the target language. The meaning of the source text is arrived at by using the source text itself as input, as well as dictionaries, encyclopedia, other texts etc. When the meaning has been obtained, it can be expressed in another language. This view of translation is very close to the interlingua view.

Even if the way to good MT is not necessarily to simulate the way in which humans do translation, MT translation theory certainly profits from human translation theory and vice versa. At the same time, we think that at least for machine translation, it is not obvious which method, transfer or interlingua, gives the best results. For the time being too little experience from interlingua system has been gathered, so future may show.

3.2. Knowledge-based translation

One of the ways in which to improve machine translation is to add more information, in particular semantic information and extra-linguistic information. The knowledge is stored in a knowledge base, or as it is called in KBMT-89 [Goodman and Nirenburg 1991], 'ontology' or 'concept lexicon'. In KBMT-89 this knowledge is used to obtain an unambiguous interlingual representation, but in fact a knowledge base of this type can also be used to augment a transfer system. The way in which a concept lexicon differs from an ordinary lexicon is that it contains representations of objects and interactions.

In KBMT, the interlingual representation is reached in a number of steps: first the syntactic analysis is performed using an LFG (Lexical Functional Grammar) like approach; then the lexical entries are translated into their interlingual counterparts using the concept dictionary, and finally structural changes from the LFG structures into interlingual structures are made if necessary.

Concept dictionaries and other types of 'world' knowledge are certainly a useful way of augmenting an MT system, and just like other modules the knowledge base can be more or less truly language-independent. Language universality or language independency is sometimes claimed, but it is well-known that this is an ideal claim, and that for practical purposes a certain amount of language dependency is acceptable, - e.g. as long as differences between the two languages in question are neutralised, the similarities need not be described in much detail.

3.3. Translation by analogy

The systems mentioned above are normally rule-based, i.e. they consist of grammar rules, lexical rules etc. More rules leads to more sophistication and more complex systems, and may in the end give systems that are quite difficult to maintain. Consequently, alternative methods have been sought. Translation by analogy is also called example-based or memory-based translation [Nagao 1984].

An analogy-based translation system has pairs of bilingual expressions stored in an example database. The source language input expression is matched against the source language examples in the database, and the best match is chosen. The system then returns the target language equivalent of this example as output, i.e. the best match is based only on the source database, different translations of the source are not taken into account. Just as for translation memories, the analogy-based translation builds on approved translations, consequently the quality of the output is expected to be high.

However, purely analogy-based systems have problems with scalability: the database becomes too large and unmanageable for systems with a realistic coverage. Consequently, a combination of the rule-based approach and the analogy-based approach is the solution. We are seeing many proposals for such hybrid solutions and this is certainly one of the areas that will bring practical MT further.

3.4. Refined linguistics

MT will profit from most of the advances made in computational linguistics, including better linguistic formalisms, better semantic theories, development of discourse theories and their implementation etc.

4. MT integrated with other technologies

4.1 Spoken language translation

One of the really interesting and challenging integrations is that of MT and spoken language systems. This is still at the research level, even if some of the research takes place in industry, but it is clearly maturing. Well-known very large projects are the Japanese ATR project and the German Verbmobil project [Kay et al 1994].

As can be seen from the research projects' choice of themes, the first commercial systems will focus upon very limited domains and dialog types, e.g. hotel reservation, travel information and reservation, etc. Automated systems may help people to communicate across language barriers, and across the world's time differences. As human speech is imperfect (false starts, repetitions, syntax errors, noise) and as speech recognition modules are imperfect, spoken language translation systems receives imperfect input and therefore is even more prone than ordinary MT to give imperfect output.

The combination of the fact that humans do not always communicate in a totally grammatical fashion in foreign languages (but still are often able to understand each other), and the fact that we have accepted that MT of written text cannot be expected to be perfect, probably leads to a certain indulgence from which the first spoken language translation systems will profit.

4.2 Other integration

In cyberspace we find and will find even more in the future, a multitude of other tools and systems with which MT can be integrated. Here, we shall just mention the combination of tools for information retrieval and abstracting with MT. This gives hitherto unknown possibilities for information retrieval. Texts can be found on the Internet, and they can be translated for information purposes by MT, probably also available on the net. Internet search engines already now provide language recognizers, so that only texts in the languages specified by the user will be selected. Consequently, multilingual information access tools will be available to the ordinary Internet user, - apart from a slight problem with the availability of languages.

5. MT integrated with other tools

One of the things that have changed because MT now runs in production is that MT is no longer seen as a stand-alone tool. MT can give a real productivity increase and job satisfaction only when it is integrated with other tools in the entire document production chain. The first requests from users focussed on layout preservation, but now a much larger range of tools can be used in the 'translator's environment' so users are not content with just layout preservation. However, here is room for improvement. This is a clear and well articulated user requirement and it will surely be fulfilled in the future.

First of all, a good interface between the word processor or document processing system and MT or other translation tools is necessary. Today, some of the desktop publishing systems seem not to communicate well with translation tools. Such systems have been thought of as stand-alone, not taking the massive need for translation into account. Such systems will either change their interface/create an interface or disappear from the market.

This papers deals with MT, but we cannot leave the subject without mentioning TM - translation memory. Translation memories have been seen as an alternative or competing technology, but here again the future lies in integration: already environments exist which integrate TM and MT the obvious way: first use TM and then use MT for the text which was not treated by the translation memory. This is a first step, but as we have seen above (3.3 Translation by analogy) there are certainly more advanced ways of integrating the two techniques and we will see practical results of research and development in this area - with the distinction between two clearly different types of system disappearing.

One of the heaviest components of an MT system is the dictionary and/or term base. It is very expensive to create terminology, and to convert it to the format necessary for a specific system. This is really a stumbling block for the application of MT systems, so future development ought to make exchange of terminology and other vocabulary easier. There are two aspects of the necessary solution to this obstacle: 1) Ideally, the coding formats for lexical entries should be harmonised for all MT systems. This is a task for MT developers and it will not be solved in the very near future. What *can* be done however, is to create conversion programmes between the different formats (the conversion can be only semi-automatic, of course, as systems require information at different levels. Investigations of this problem are taking place i.a. in the EU project OTELO with the participation of some of the big players on

the MT market. 2) The other aspect is the proprietary nature of terminology. Companies who have developed a terminology for their market segment are very often not prepared to share it with others. If this problem is not overcome, conversion programs and harmonization make very little difference. Since dictionary development is one of the most expensive tasks in MT development and maintenance, both aspects of this problem will have to be solved.

6. Organising the work with translation tools

Translation tools should be introduced in a translation department after thorough preparation. The savings quoted above will only materialise, if the staff taking care of the translation process organises itself properly. First step is a workflow analysis of the steps of the translation process in the new environment. This sometimes means inventing tasks that did not exist before or that were almost invisible. E.g. how does the document come from the author to the desk of the translator? What steps does this include? What types of information is needed, in what format? The time used for each task should be measured, and tools should be developed for the most time-consuming tasks if possible.

The introduction of tools has also shed light on the fact that not all of the preparatory actions for a translation have to be performed by a highly educated translator. If again we turn to the European Commission, the SdT have just implemented their Euramis workbench for translation of documents. It is interesting to see that this is not a workbench for *translators,* but rather a workbench for *translation.* The introduction of the Euramis platform has helped identify or formulate tasks which can be performed by secretaries. In some cases the introduction of slightly modified procedures may enable secretarial staff to perform tasks that were otherwise assigned to translators. This gives more job satisfaction at both ends.

7. Conclusion

We have seen that technology is available although there is room for improvement. We have also seen that extremely interesting possibilities lie in the integration of various techniques, and that users are becoming a strong player on the scene. This all points in a very positive direction and I am sure we will meet again in the future, also in the far future, although we may not call our meeting MT Summit then, if MT has been so much integrated with other technologies that it can no longer be distinguished.

However, before stopping, I would like to point out two areas where improvement is needed. First, far from all those who could benefit from MT and other language tools are actually using it; this is mostly because of lack of information and it will change because of market requirements, because of campaigns etc. so probably this problem is not so serious.

The other problem is the language problem: the brilliant future is there for the so-called major languages, but users of less spoken languages need MT and other tools just as much or even more than users of English and French. For some languages the market is not sufficiently large, which means that users of such a language will lack the tools which are otherwise available. This lack of tools will have an obvious economic effect, but also a cultural effect by excluding some languages from participating in an otherwise flourishing multilinguality.

As the problem cannot be solved by the market forces, governments and funding agencies have to take a quite active role in the protection and reinforcement of such languages. So my last words in this talk about the future of MT are a plea to those responsible for the protection and reinforcement of less spoken languages: Be careful to ensure that your language participates in the future of MT!

Acknowledgements

Thanks for valuable input goes to J. Hutchins, C. Brace, V. Hansen and D. Theologitis, all of the EAMT Committee.

References

Boitet, C. *Factors for success (and failure) in Machine Translation - some lessons of the first 50 years of R&D,* In: MT Summit V Proceedings, Luxembourg 1995.

Carter, D. and R. Becket, M. Rayner, R. Eklund, C. MacDermid, M. Wirén, S. Kirchmeier-Andersen, C. Philp *Translation Methodology in the Spoken Language Translator: An Evaluation,* In: Spoken Language Translation, Proceedings, ACL/ELSNET, Madrid, 1997, 73-81.

Goodman, K. and S. Nirenburg (ed.) *The KBMT-89 Project: A Case Study in Knowledge-Based Machine Translation,* Morgan-Kaufmann, San Mateo, 1991.

Horiguchi, K. and A. Franz A *Formal Basis for Spoken Language Translation by Analogy,* In: Spoken Language Translation, Proceedings, ACL/ELSNET, Madrid, 1997, p. 32-39.

Kay, M. and J.M. Gawron, P. Norvig *Verbmobil: A Translation System for Face-to- Face Dialog,* CSLI Lecture Notes No. 33, Stanford, 1994.

Maegaard, B. and V. Hansen *PaTrans, Machine Translation of Patent Texts, From Research to Practical Application,* In: Second Language Engineering Convention, Convention Digest, London, 1995, p. 1-8.

Melby, A. *Why we need Data Exchange Standards,* In: The 1997 LISA Tools Workshop Guidebook, Geneva 1997, p. 9/1-9/8.

Nagao, M. *A framework of a Machine Translation between Japanese and English by analogy principle* In: Elithorn and Banerji (ed) Artificial and Human Intelligence, North Holland, 1984, p 173-180.

Nida, E.A. *Towards a Science of Translating,* Brill, Leiden, 1964.

Reiss, K. *Texttyp und Übersetzungsmethode,* Heidelberg, 1976.

Reiss, K. and H. Vermeer *Grundlegung der allgemeinen Translationstheorie,* Tubingen, 1984.

Spoken Language Translation, Proceedings, ACL/ELSNET, Madrid, 1997.

The 1997 LISA Tools Workshop Guidebook, LISA, Geneva 1997.

Theologitis, D. *Integrating Advanced Translation Technology, Keynote Address.* In: The 1997 LISA Tools Workshop Guidebook, Geneva 1997, p. 1/1-1/35.

Tsujii, J. *Future Directions of Machine Translation,* In: COLING '86 Proceedings, Bonn, 1986, p. 655-668.