MT FROM AN EVERYDAY USER'S POINT OF VIEW

Annelise Bech
Lingtech A/S
Vesterbrogade24
DK-1620 Copenhagen V
Denmark

Email: lingtech@login.dknet.dk
Phone:+45 3325 7171
Fax: +45 3325 6171

Abstract

This paper discusses the experiences of the specialised Danish translation company Lingtech in its use of MT for the translation of technical texts. The background and motivation for setting up Lingtech as an MT-based company is outlined. After a short general presentation of the PaTrans MT-system, the different tasks we have to perform in relation to our use of MT and the way this work is organized in order to achieve maximum cost-efficiency are described. This leads on to the discussion of problem areas for the everyday user in terms of ergonomy and tools for what may be called 'peripheral' tasks, e.g. pre- and post-editing texts, and dictionary maintenance. In the course of gaining experience in running an MT-based organization, we have identified crucial areas, where even relatively simple tools can have quite an impact on the overall productivity and profitability of using MT. Given the state-of-the-art within language technology many useful tools can now be made for the MT-user; however, we argue that too little attention has been given to these aspects so far and that they may indeed be critical to the commercial success of machine translation.

0. Introduction

Even though machine translation has been around for years, it still seems that only few private companies have exploited its possibilities to any greater extent, let alone based their translation tasks on the use of some of the more advanced solutions. The Danish translation company Lingtech, situated in the center of Copenhagen in Denmark, is a clear exception to this.

Since the end of 1993 when the machine translation system was first introduced into the organization, PaTrans has been used to translate technical texts from English into Danish. This has lead to increased productivity, better consistency and quality in the finalized translations, and to a considerable reduction in translation costs.

To us it has become clear that using machine translation in its proper context pays off. However, we have also learned that the gains can be even more impressive when more attention is devoted to work-flow organization and to providing still better ‘peripheral’ tools to help MT-users tackle some of their time-consuming, costly, and tiresome daily tasks. Experiences from Lingtech's past and present emphasizes the paramount importance of these aspects for the true commercial success of machine translation. This paper illustrates the point. But first, how did it all start?

1. Background and Actors

As for Lingtech, it all goes back to the introduction of the European Patents Agreement. With this agreement, it was predicted that the number of European patents to be validated in Denmark would increase dramatically. For a Euro-patent to be validated, the legal requirement is that a translation of the original text into the language of the designated country be performed. Hence, a need for an increase in translation resources was rightly foreseen within the Danish patent business.

Forward-looking leaders from the two Danish patent attorney companies Hofman-Bang & Boutard A/S and Lehmann & Ree A/S planned to meet the increased translation demand by making use of modern technology. They embarked on a joint venture in setting up a new company, i.e. Lingtech, a dedicated translation service envisaged to rely extensively on the exploitation of advanced machine translation technology for the bulk of its translations. The goal was to have a company that was at the competitive edge with respect to both translation capacity, quality, and translation costs. Lingtech was to be the implementation of the 'high-tech translation factory'.

At the time of the conception of the translation factory back in 1990, no existing machine translation system fulfilled the requirements of being able to support a continuous process of feeding in documents in English and producing high-quality translations thereof in Danish. Therefore the companies established a cooperation with the Danish Centre for Language Technology in Copenhagen, for the development of their own purpose-designed system.

Founded in the linguistics and machine translation technology of the European Commission's Eurotra project, the Centre for Language Technology developed the PaTrans MT-system and delivered it to Lingtech near the end of 1993. (For more details about the development project, see e.g. Bech 1992; Bech 1994; Maegaard and Hansen 1995)

After a period of testing and improving the system in the actual production environment and building up basic technical dictionaries for the subject fields of the texts to be processed, machine translating texts became part of the daily life of Lingtech and has become increasingly so ever since.

Before we go on to the actual translation work-flow at Lingtech and the tasks and issues specifically related to machine translation, a few words about the overall design of the PaTrans translation system and the kind of texts we translate are in place in order to set a proper context.

2. PaTrans and Patent Texts

PaTrans, which is short for patent translation, is a transfer-based system the basic idea of which is to let the texts for translation undergo thorough linguistic processing, both lexically and grammatically. At the heart of the system is a comprehensive grammar for English (analysis) and Danish (synthesis), respectively, both specifically designed for and tuned to the patent text type. The system operates in batch mode and runs under Unix on workstations.

Lexically, the system distinguishes between general words and domain specific technical terms, and has separate dictionaries for holding these. The general words are items from the closed word classes and items from the open word classes appearing in a patent text irrespective of the subject field it concerns, i.e. verbs, adjectives and nouns such as 'describe', 'present', and 'invention' reside in the general dictionary. Contrary to the general dictionary, which is not modifiable by Lingtech, the domain specific dictionaries, in short called term dictionaries, are created and maintained by us. We can define and have as many different term dictionaries as we find suitable.

Term dictionaries hold the technical words and expressions which have a particular meaning and translation within a given subject field. Thus for example, 'component' is in our chemical term dictionary with the Danish translation 'bestanddel' and in our mechanical term dictionary with the Danish translation 'komponent' reflecting the different meanings of this word within different fields. The set and order of dictionaries to be used for the translation of each individual text is determined by the system operators at Lingtech.

Up till now, we have concentrated on machine translating patents. A patent text is a legal document the purpose of which is to describe a new invention and to characterize it and its innovative aspects in relation to other similar or related products or processes in order to protect its holder. Both the descriptive and the legal nature of the text type is clearly reflected in the language of the texts, which grammatically spans from elaborate narrative to very complex and compact linguistic constructions. Lexically, the texts are characterized by a very high percentage of their vocabulary being terms or technical expressions for which a very specific translation is required within a given field. The topics dealt with in the texts range from lubricating oil compositions, diapers, body-shapers, and toys to shampoo, just to mention a few examples.

3. The 'High-Tech Translation Factory' in Operation: Work-Flow and Tasks

The volume of patent texts translated by Lingtech currently amounts to some 8-9 million words per year or in the order of 35.000 pages. A rough 75 percent of the total volume is texts to be translated from English into Danish, and hence candidate for machine translation. Texts in other languages are translated manually by freelance staff.

Incoming texts for translation are registered in Lingtech's administrative system and then assigned to a translator. Patent texts to be translated from English into Danish are reviewed for their suitability for machine translation according to our experiences with the performance of the MT-system and the cost-efficiency of previously translated texts. We will return to a discussion of the selection strategy itself, as it has proven to have quite an impact on our results.

When texts are accepted for machine translation, we need to perform several of what may be called 'peripheral tasks', i.e. preparing texts for the system, identifying and coding new technical words and expressions in the term dictionaries, and finalizing (post-editing) the machine translated texts.

The preparation of the texts and the coding of the new entries in the dictionaries are taken care of by our computational linguists, who also operate the translation system. The translation itself is performed by the system without any interaction from the user. The finalizing is handled by technical experts in the relevant subject fields of the texts.

Given the fact that we currently see a rough 50 percent saving in costs per machine translated word compared to a manually translated word, the goal is obviously to maximise the number of words translated by machine. Thus, we are constantly on the look out for new ways to further optimize by way of introducing additional tools or reengineering our work process. Over the past one and a half years, we have obtained remarkable improvements by relatively simple measures and little effort. With the same human effort we have gone from machine translating some 1.2 million words per year to the present more than 3 million!

In introducing machine translation in the first place, we went for higher productivity, consistency and lower costs; and to us as a commercial company the cost-efficiency factor is obviously still important. The improvements we have seen so far are the results of seriously addressing what may seem simple and trivial tasks. Our strategy has been - and still is - to aim for the highest degree of automation of peripheral and ancillary tasks in order to reduce the human effort necessary for carrying them out.

In the following discussion, we will focus on some of the areas we have worked with. Even though we are pleased with the results of our efforts, there are certainly still outstanding issues and room for improvements.

4. The Factory Manager's Credo: Work Smarter

Using an MT-system can involve a lot of hard and tedious work if we do not some times stop and critically review tasks, tools and procedures. Focusing on the more time-consuming and repetitive tasks which require human resources is where to begin if we want to get from working hard to working smart.

4.1 Hard Finding - Easy Coding

As part of the preparation for translation, the text needs to be checked for the occurrence of terms or expressions to be added to the dictionary, i.e. the new entries to be coded. This is done by running the system in dictionary look-up mode, the result of which is an alphabetised checklist of all the words in the text with indication of their status as 'known' or 'unknown'.

For a long time this checklist was the only basis for the further work with expanding the term dictionaries of the system. And this proved to be problematic for several reasons; first and foremost because the text itself had to be consulted extremely often during this preparatory task, something which is both cumbersome and time-consuming. In practise the checklist facility was an insufficient help in the actual production environment. The more critical problem area concerned multi-word terms.

Multi-word terms have to be in the dictionary as one entry for the system to recognise and treat them correctly. For new multi-word terms to be coded, the checklist only provided little help as it had nothing to say about what the new multi-word terms are. Because the general definition of a lexical unit in the system is a string of characters delimited by blanks, only the components of what should be a multi-word term will obviously be found on the checklist. Therefore it was necessary to manually go through the text in order to identify multi-word terms. For example, unless the multi-word term 'lubricating oil composition' (DA: 'smøreoliesammensætning') has been coded as an entry in the term dictionary, the three component parts will appear individually on the alphabetised checklist, i.e. 'lubricating', 'oil', and 'composition'. However, when a multi-word term has been coded in the dictionary as one entry ('lubricating_oil_composition'), the system will correctly identify it as such, when processing a text in which it occurs. That is, a 'known' multi-word term will also appear on the alphabetised checklist.

Complicating the matter further is the fact that compounding is very productive, i.e. even when the dictionary contains 'lubricating_oil_composition', this unit may also occur in a text as part of 'lubrication oil composition container'. Consequently, this multi-word term also needs to be coded as an entry in the dictionary.

In our text type there is a very high number of multi-word terms and the process of identifying them was - as can be gathered - quite time-consuming and rather tedious. To speed up this work an interactive concordance facility was added to the pre-editing tool, so that it was no longer necessary to read through the text. This additional tool was well-received and found to be useful. It improved preparation productivity and enhanced the ergonomy for the staff.

However, why not have a fully automatic tool that could scan through the text and propose new candidate multi-word terms. Confident that such a tool could be made, we specified our requirements and had our system developer program it for us and integrate it into the pre-editing environment. We can now specify rules for candidate compounds and multi-word expressions, and found candidates are highlighted in the text in the pre-editor. The role of the alphabetised checklist has consequently been drastically reduced. The speed with which new entries to be made in the dictionary are identified has now at least doubled. The coding of the entries, once we know what they should be, is an easy task using the dictionary coding tool of the system.

4.2 Butterfly-Like Diapers and Sugar-Coated Nuts

Originally the idea was to have a large number of highly specialised term dictionaries. For each text to be machine translated, the relevant set of dictionaries to be used must be specified and in a prioritised order to solve lexical ambiguity by having the system prefer entries in the dictionary with higher priority. In practice this has turned out to be cumbersome.

First of all, there is no straightforward relationship between the subject field of a patent text and the lexical items that may appear in it, e.g. a diaper may have a shape like a butterfly! Secondly, also specifying the order of the dictionaries for the translation soon turned out to be troublesome.

As the dictionaries grew in number and size, it obviously became hard to keep track of which dictionary contained what entry. Consequently, the translation process created interesting new products such as a cereal with 'sugar-coated nuts' [where 'nut' was translated as in its 'nuts and

bolts'-sense]. More often though, the result was that a component of a multi-word term had been found in a dictionary with higher priority, causing the system to ignore the entry for the multi-word term in a 'lower' dictionary.

As a consequence, the term dictionaries now cover less narrow fields; we have two major ones split according to the general patent distinction of chemistry and mechanics, and three minor dictionaries for the terms, which do not obviously belong in either of the former two.

Having reduced the problems somewhat this way, the question of in which dictionary to place a given entry still occurs regularly. Here another recent improvement to our pre-editing environment helps us out. After having run a text in dictionary look-up mode, the user can request to see the text displayed with 'known' terms in different colours depending on which dictionary they reside in. As it is also possible to have the coded translation for the term shown at this point, we now have a far better situation than before where the different dictionaries had to be consulted in order to look up an entry. The staff performing the preparation report that their working situation has greatly improved, and we expect to see productivity increase.

Also we have had a facility added to the system's dictionary component that allows us to collapse two or more term dictionaries into one for a translation rather than having to order them according to priority.

4.3 Words, Words, Words... - Selecting the Right Texts for Translation

Since the ratio of terms and technical expressions in patent texts is very high, they are well suited for machine translation. However, given that we need to have quite large dictionaries due to the nature of patent texts that abound with specialised terminology from within the various fields, the speed of expanding the dictionaries is naturally a critical economic factor for us.

As we have seen, the critical aspect is not so much the coding of new entries; it is the identification of what to code, which takes up resources. Even with term dictionaries currently holding some 80,000 entries in total, some texts require a coding effort where the ratio between the total number of words in the text and the new entries to be made is too high from an economical point of view. As the distribution of subjects our texts concern is out of our control and not very predictable, we have had to develop pragmatic strategies to prevent us from getting bogged down by dictionary work.

Consequently, we have improved our strategy for selecting texts for machine translation by taking into account the length of the text and our impression of what the coding ratio is likely to be. Currently the selection is made by humans using their built-up experience and their knowledge of the coverage of the term dictionaries. It would, of course, be far more ideal to have some tool, which could automatically process the texts and report back a score on the basis of which the optimal selection could be made.

With our present strategy, we have achieved a situation where we 'only' add on the order of two percent new entries out of the total amount of words translated on the system. For patent texts, it is an open question how much further down the average percentage will ever get, though.

4.4 Easing the Burden of the Post-Editor and Improving the Performance of the System

Even advanced MT-systems can at their best only produce raw translations that need to be post-edited and finalized. Therefore the importance of providing efficient tools for the post-editing staff has long since been established. However, there are also proactive ways to help ease the burden of the post-editor and to improve cost-efficiency: In the selection of texts to be machine translated the texts yielding the better output should be preferred and the performance of the system must continuously be improved.

As for the selection of texts, we again have to base this on our intuition and past experience. We pretty well know which kinds of texts to reject, and we have accumulated knowledge that enables us to make good guesses. But some times we are in for nasty surprises. A tool for checking candidate texts for machine translation for the degree of compatibility with the grammatical coverage and performance of the MT-system would be a great help, we believe.

Going on to the question of quality in the performance of an MT-system, neither developer nor user would probably disagree that the highest level possible is to be striven for. However, their understanding of high quality may differ radically. From the technical point of view the quality of a system tends to be rated predominantly on the basis of the number and the complexity of the grammatical constructions it can handle and translate correctly; this holds true for the user only in part, once above a certain base-line, of course.

In the, let us call it advanced improvement phase of an MT-system, it is easy - both for users and developers - to fall into the trap of only looking at phenomena in isolation, without considering the full context and taking the post-editing task sufficiently into account. There are many good reasons for this, the main one probably being the simplicity of registering, testing and checking. It is 'easy' to deal with how for example imperative constructions are dealt with in the translation process and appears in the output.

We have also been wound up in this way of looking at things when reporting back to our system developer our wishes for corrections and improvements to our MT-system. However, it has become increasingly clear to us that it may be a dead-end street. We have therefore recently begun trying to focus on the impact and importance a certain improvement may or may not have on the global post-editing task. The cost-efficiency factor is naturally at play here again, too.

In structured questionnaires to the post-editing staff, we have asked for feedback with respect to their opinion of the frequency and view of various types of faulty output from the system. Based on this feedback, we try to determine and prioritize the improvements to be made to the system. We have so to speak given the major role to the ‘frequency and irritation factor’.

The strategy is not unproblematic, however, because it relies on ‘general impressions’. We know of course that the irritation factor is per se subjective and individual; but the recording of the frequency of occurrence of an error type is not; it can be quantified. Doing this manually is a costly and time-consuming enterprise, what we could wish for is some kind of tool, which could automatically analyse quite huge corpora of unedited output material from the MT-system for specified flaws and report back findings in a clever way.

5. Conclusion

When striving for the stars, the chance is that you may get to the tree top! At Lingtech we have been able to more than double the number of words machine translated without adding more human resources. Naturally the human effort machine translation requires, decreases as the dictionaries grow in coverage and size. However, as we have seen, the implementation of sound work procedures and the provision of efficient and clever tools for peripheral tasks have an enormous impact on productivity, ergonomy, and the cost-efficiency of the whole enterprise.

For the user as well as the MT-developer, it is naturally a process of learning by experience and practise where the problem areas are and what can be done to them. However, we claim that aspects such as those illustrated in this paper and others related to the peripheral tasks to be performed by the user have received far too little serious attention up to now.

It is in the production environment that machine translation has to demonstrate its viability. In its proper context machine translation pays off, yet the true commercial success of it, we believe, now depends on inventing, designing, and providing more clever and efficient tools for the user. Technologically it is possible - what is needed is perhaps a change in status of such tools. Research in and the development of ancillary tools need to become prestigious.

A short Postscript

Credit is due to Mr. Viggo Hansen, the former director of Lingtech, now managing director of Hofman-Bang, for starting up and providing the basis for the further development and application of MT at Lingtech. The present author took over the responsibility for running and expanding Lingtech's activities after Mr. Hansen in late 1995. Having previously in her career worked with research and development of MT-systems, being put at the other end as user of MT has been an interesting experience. Former views and priorities get set into a quite different perspective with the result of some of them turning upside down and completely new ones presenting themselves.

References

Bech, A. (forthcoming). Using Language Technology - Does it Pay Off? In Proceedings from the XXII International Association Language and Business Conference 1996.

Bech, A. 1994. Kan PaTrans oversætte nu? In *Bits & Bytes: Datalingvistisk Årsmode nr. 3.* Odense Universitet, Denmark, 69-73.

Bech, A. 1992. PaTrans-projektet. In *Skriften på skærmen, no. 6.* Handelshøjskolen in Aarhus, Denmark, 107-117.

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