[*Terminologie et Traduction*, no.1, 1986]

MACHINE TRANSLATION IN THE USAF

Dale Bostad
USAF Foreign Technology Division, Dayton, Ohio.

Introduction

Machine translation is used by the USAF to support studies carried out by scientific and technical researchers who need to stay abreast of foreign developments in a wide variety of technical fields. Most of the translations produced come from open-source literature. A researcher has a broad base of information at his disposal, including extracts, abstracts, cover-to-cover translations of foreign journals produced by publishing houses, and other studies in the field. Machine translations are simply another source of information. It must be emphasized that the researchers - engineers, physicists, mathematicians, and computer scientists - are experts in their fields and have a wealth of knowledge about developments in the areas they are interested in.

It is difficult to define exactly how each researcher uses translations since there are several hundred users, each with his own method of research and sources of information. We do know that between 50,000 and 60,000 Russian pages are translated per year by the USAF Russian system and several thousand by the French system, and that some studies are completely based on translations and others are not, and that the amount of requests for translations varies considerably, some users requesting less than a hundred pages per year while other users request several thousands of pages.

It is the purpose of this paper to analyze the use of machine translation in the USAF, to describe the product, and to indicate future developments in the area of rapid information acquisition.

Machine translation - its use

Machine translation is used to translate large amounts of text material for rapid acquisition of content. The translations are focused in the sense that there is some determination of the relevance of the material translated. MT translations simply are not done willy-nilly on large volumes of indiscriminate material.

There is a very careful screening process using extracts, abstracts, translated tables of contents, and oral prescreening of documents. Often only limited sections of large documents are earmarked for translation. Given the selectivity of the translations, the fact that the translations are targeted for a limited audience, and the translations make no pretense to be of publication quality, the term focused translation for information content seems to best describe the process.

Some books and documents translated contain a wealth of substantive information and, in fact, it can be said that most of what is in the book is important. These types of documents are often translated cover-to-cover and serve as long-term reference for the researcher. In other cases no more than 10-15% of a document contains relevant information. And despite all indications from oral prescreening or abstracts, it sometimes happens that a translated book does not live up to expectations and the researcher makes little use of it. Simply stated, in many cases the value of a given document cannot be ascertained prior to the actual translation.

The critical point to remember is that the MT translation, although used for broad information acquisition, is the finished translation that the researcher gets. It was thus necessary to develop a translation product that could be turned out quickly, knowing that much of the translation may be of little value, and at the same time produce a translation that would be of high enough quality to give the researcher technically accurate information in those substantive parts in which he is interested. The product that evolved is what we call partially-edited machine translation.

Partially-edited MT

The USAF's standard translation product, partially-edited machine translation, meets the standards of adequacy for the users of the product. It has gained wide user acceptance. It provides technically accurate translations in a wide range of technical disciplines and is readily comprehensible to a subject-area specialist.

Moreover, it provides very rapid access to information - the paramount consideration of researchers who have to meet deadlines for completion of studies. Lastly, and this is of more concern to management than to the researchers, statistics bear out that machine translation is the most economical way of producing high-volume translations.

A partially-edited translation is produced by scrolling through the entire translated text on a video-display terminal. However, only segments of the text are actually post-edited. Not every sentence is looked at; in fact, it is our estimate that only about 20% of a given text is edited. What is to be edited is determined by a software program called EDITSYS. EDITSYS is currently used for the Russian-source system and it will soon be implemented for the French-source system. The functioning of this program will be examined in some detail.

EDITSYS

EDITSYS is a program called at the end of the translation procedure that serves to direct the post-editing, i.e. it tells the editor exactly what to look at and edit. The program identifies trouble areas in the system that need review and intercepts these conditions. As stated, large chunks of text go through unscrutinized without editing. This means that we rely heavily on the efficacy of the linguistic algorithms and our large dictionaries. To write a program like EDITSYS you have to have considerable knowledge about the strengths and weaknesses of the system and the programming expertise to highlight the weaknesses for review.

The program itself is a module that allows us to go in and test at the bit/byte level the final analysis area of sentences. Virtually all of the linguistic macros in the system can be used for testing. When a given test condition is met the program generates a full-width line of a certain character in front of the condition and this line is interspersed in the text and displayed on the screen. As an editor scrolls through the translated text he halts whenever a flag line appears and makes an editing decision. If no editing is required he erases the flag line with two keystrokes; otherwise, he corrects the error.

Post-editing is limited to the immediate environment around the flag. A skilled editor can edit 15-20 pages of Russian-English translation per hour using this technique.

Flags are generated by EDITSYS to check the following situations:

1) Not-found words. All legitimate not-found words or words incorrectly input are flagged. True not-found words are now relatively rare, since the dictionaries contain 200,000 individual entries.

2) Acronyms. All acronyms are checked to see if their expansions are correct. Thousands of acronyms are expanded in the dictionaries, but those of three characters or less require close scrutiny.

3) Rearrangement. Byte 144 indicating rearrangement is flagged. Approximately 20% of sentences from Russian are rearranged with an accuracy rate of 90%. One sentence out of ten must be edited when words or phrases are moved into incorrect slots.

4) Contiguous slashed entries. There are several thousand slashed entries in the Russian-English system and when slashed words in English occur next to each other, smooth reading of the text is impeded. The most frequent occurrences are adjective + adjective, adjective + noun, and noun + noun.

5) Spurious "good" terms. These are words that have been typed in or scanned in incorrectly but which match up against the dictionary. Examples are BOLE instead of BOLEE, S04 instead of SL04, and BIT6 instead of BYT6.

6) Uncertainty code. Byte 57,04 is tested. This uncertainty code is turned on in certain homograph routines at the point where the logic becomes tenuous, there is no statistical evidence for one dictionary default over another, and in fact resolution is a toss-up.

7) Problem words. There is a flag generated for certain problem words (about 40 in number) which the system has not been able to resolve with sufficient accuracy. This category is fluid; as routines or expressions are developed for these words they are no longer flagged. Of course, new conditions or words also arise which require flagging.

This then is what we call partially-edited MT. It is somewhere between raw MT and completely edited MT. The product has gained high acceptance, but any researcher who wants a closer translation of critical material can bring his partially-edited translation back for heavier editing. This happens for about 1% of the MT translations we produce.

The future - interactive raw MT

The USAF is sponsoring the development of interactive raw machine translation for researchers. It will feature researchers running translation procedures at their own terminals without the mediation of any translators. The Russian system is projected to be on-line in March, the French and German systems by mid-summer. This development will allow researchers immediate translations of titles, tables of contents, sentences, and even paragraphs. All that the researcher or his secretary will have to do is to type in the material to be translated on a terminal, make some menu selections, and the raw MT will be displayed on his screen in a matter of seconds. This type of interactive translation can be called information scanning in its purest form. Interactive raw machine translation, of course, should not be confused with interactive translation systems like Weidner or Alps.

It is expected that this use of MT will replace oral translation screening and serve as a mechanism to quite precisely pinpoint segments of documents and books to be translated. It will also serve as an impetus to get researchers more involved in MT, and give them more control in obtaining instantaneous information without going through the bureaucracy. Interactive raw MT should be especially valuable for researchers who have no knowledge of Russian, French, or German.

Programming considerations

The current procedures of Systran-based systems must be fundamentally changed to support interactive raw MT. The approach can be briefly described.

Systran-based systems presently run in a batch environment designed for the efficient processing of large volumes of text. For example, low-frequency words are sorted for lookup in the main dictionary so that only one random access search is needed for identical words. FTD often runs 10,000 sentences at a time so it is necessary to keep the random access to a minimum. But in small-volume texts, sorting of low-frequency words for dictionary lookup and then resorting them back to the original sentence requires more overhead than random access searching.

The approach is to modify the Systran-based systems by removing the two sorts and thereby reduce the current five-step procedure to a single step. Simultaneously the systems will be moved over from OS to run under CMS (Conversational Monitor System). The system will be loaded on a minidisk on the host and all users will have their own virtual machine running on the host.

All that will be necessary is to link to the minidisk, access the translation procedure, execute a CMS run command, and the input file will be translated and returned to the user virtually immediately where it will be displayed on his terminal or it will go to a disk file where it can be accessed and displayed at any time. The most probable scenario is that CMS files would be created directly on the host computer. However, input files could also be created in local CMS sessions and sent to the host and brought down from the host using VM/PC Serve. DOS files created on intelligent terminals could also be used. In any case, as many as 1,000 terminals will be available to users to access the translation procedures.

An executive procedure will be written to run under CMS to guide the analyst through a menu of options that will include language selection and automatic dictionary allocation to the file being created. The EDITSYS program will not be called during this type of translation.

Material for translation will be input on personal computers using transliteration conventions already in place. There should be few problems for researchers and secretaries in picking up the these conventions. Indeed, many researchers have some working knowledge of Russian and this should make it even easier.

In preliminary discussions researchers have been very receptive to the idea of automatic on-line raw MT. The concept of on-line automatic raw MT translation will only work with highly developed systems with very large dictionaries. Researchers will not use the systems if there are junk translations and not-found words. Hence, Systran-based systems are ideally configured for this type of translation.