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

THE IMPACT OF CONTROLLED ENGLISH ON MACHINE TRANSLATION

J. Richard Ruffino & Peter F. Demauro,
Xerox Corporation, Webster, N.Y., U.S.A.

(This paper is adapted from a presentation given at the American Translators Association Conference in Miami, Florida, in October 1985. It has been published in the proceedings of that conference by Learned Information, Inc., Medford, New Jersey, 1985.)

Introduction

The most fascinating aspect of working with machine translation is the study of how the computer interprets data in comparison to the human mind. Those who first conceptualized translation from one language to another using the computer approached language primarily as a science, to an extent overlooking the fluid and affective aspects of human expression. The task of the machine in producing translation is to emulate the human mind. Although systems vary in their level of sophistication and therefore the degree of success achieved at this task, the underlying theme of all translation systems is to create an "artificial intelligence" that can interpret language. The first developers of machine translation failed to realize the task that they were up against, because the view of language as a science gives the impression that although complex, the phenomena of speech and expression can be harnessed, defined and contained. Such a view would then lead one to believe that natural language is in itself "controlled" and therefore a candidate for computation and for the computer. Starting with this basic premise, we can see how hopeful one can become in trying to establish a system for language translation. This optimism, however, fades as we move away from language as a science to language as a human art. This is not to say that machine translation is without value. As an aid to the technical translator it is a remarkable tool, but by understanding its limitations we can take greatest advantage of its potential.

Hierarchy of control

In the examination of English as the source language in a machine translation system, it becomes evident that there are different areas of language input that can be "saturated" and/or "controlled". Some definition of what we mean by these terms is appropriate. "Saturation" means that a particular consideration or aspect of the language can be saturated by the dictionary or programs used in the translation process. For example, every word in Webster's or the Oxford dictionaries could be entered in the dictionary used for computer translation. Although this is a formidable task, it is possible to expand a database to include every word we use. At first thought, such an endeavor seems to offer great hope for the creation of a very productive translation system because we are able to define our range and eventually meet the target. In addition to being able to provide a translation for every word, those who are not experienced with translation would most likely be under the impression that a word for word exchange will yield an acceptable translation.

Therefore, in the beginning, hope for high quality machine translation ran high from the standpoint of word for word lookup and the quality of translation it could provide. What was ignored was the fact that even though it is possible to enter every word in a machine dictionary, it is not advisable, mainly due to cost and necessity. Consequently, even the first level in our hierarchy of difficulty of natural language input requires some control, that is, a restriction of the vocabulary that can be used in the input language. By "control" we mean that an aspect of natural language can be limited or restricted without seriously altering the language as perceived by the native speaker. In this first example, we see that we can saturate the area of vocabulary by adding every word to our translation dictionary, but it is preferable to control the addition of words and this will not adversely affect the input language quality.

Word combinations

The next step in our hierarchy of control after single words is combinations of words that require special meanings. In looking at natural language, we see that on the first plateau of vocabulary, saturation is possible, but not advisable. Control of the input is the answer that will give us more productivity in the output translation. The next level is that of combinations of words in both nominal and verbal phrases. In general, verbal phrases are of a higher order than nominal phrases because of greater inflection, and greater possibility of divergent meaning. Consider "run into" or "put on" as examples of this. Both nominal and verbal phrases can be saturated. That is, it would be possible to enter every noun phrase (combinations of nouns and nouns or adjectives and nouns that name a unit or independent entity) and verbal phrase (combinations of verbs and adverbs or prepositions that alter the meaning of the basic verb). As in the case of basic words, this is possible, but even more difficult as there exists nowhere a listing of every nominal and verbal phrase in our language. Saturation on this second level of the hierarchy becomes almost impossible. Control as well becomes more difficult because now where we have permitted the use of a basic word, we are obliged to allow its use in combination with other words, regardless of how variant the meaning is from the basic components. The first step in control is pinpointing the nominal and verbal phrases used in a particular source language. The most expedient means of doing this is through a process called "pre-edit". Pre-edit is a review of machine translated text before the scheduled time of translation for production in order to capture any expressions which will have to be entered in the dictionary. For example, the verb "take" and the adverb "out" will both obviously be included in the basic dictionary, therefore, it is very difficult to restrict the combinations of words regardless of how we much we might believe that it would help us to gain the necessary control over our input text. Through the process of pre-edit we can at least become aware of the fact that the verb phrase "take out" is being used and how it is being used and can enter it in our dictionary with the appropriate meaning. Pre-edit then is a tool not so much for "tweaking" texts to make them translate, but rather as a method of control which allows us to saturate our translation dictionary with the nominal and verbal phrases that are being used in a specific text.

Meaning and part-of-speech multiplicity

Moving up to the next level of difficulty of control we come to the problem of homographs. For the purposes of machine translation, homographs have been primarily considered from the standpoint of part-of-speech, that is, a word which has multiple and what we should call "active" parts-of-speech. I include the word "active", because so many words have multiple parts-of-speech in the English language, but many are exaggerated and of such a low frequency of use that we should not really bother with them. How many times a day do we observe "the wonderful manage of the thoroughbred?" (as a noun, manage is the action and paces of a trained riding horse). Computational linguists have rightfully always given priority to the homograph as part-of-speech because this is the most common and has the most impact on the quality of translation, as homograph resolution is often the deciding factor on sentence parsing and whether the sentence will make any sense in the target language. It is very difficult to control the use of homographs, however, because of how widespread they are in our language. Pre-edit can only tell us that a homograph is being used, so that we can be sure to enter it in our translation dictionary. Pre-edit does not help us gain the correct resolution of the homograph in the translation unless we tamper with the input text by inserting articles (clean surface, the clean surface or clean the surface?), rewriting the text or substituting the homograph for another word. Although we can saturate our translation dictionaries with all active homographs and we can control the input text by reviewing and rewriting, this type of control is not recommended because it is very time-consuming and costly and therefore, negates the productivity to be gained by using computer assisted translation in the first place. The problem of multiplicity of meaning presents an even greater challenge, although examples of it are not as numerous as part-of-speech homographs. When a single word has a variety of meanings, thus requiring a variety of translations based on the English usage, there are some things that we can do, but no solution is foolproof. Take for example the word "file". This is a noun/verb homograph and it is also a homograph of meaning to be interpreted as an "archive" or a "tool". It is possible to define certain parameters of usage to add nominal and verbal phrases and relationships to our dictionaries and thus assign meanings based on these contexts. For example, every time the word "file" appears in the context of "file server" we can assign the correct meaning of "archive". Likewise, if the verb "file" takes "report" as its object, then we would be relatively safe in assigning "to archive" as the meaning of the verb in this case. Such delineations, however, do not provide a complete answer for machine translation as words often appear in isolation without contextual clues that can serve as indicators to the machine programs. Another approach to divergent meanings is to assign meanings based on the topic of the English text. Logically, if we are translating a manicurist's training course, "file" will most likely be translated as a "tool". But we have seen that even in highly technical texts such assumptions are no guarantee and that words of multiple meaning are often used in their different senses within the same text. With the homograph of part-of-speech we begin to lose even the possibility of control of our input text and there can be no saturation of our dictionaries when it comes to homographs of meaning.

Idiomatic and colloquial expressions

This next level in our discussion is really composed of two areas which we will consider together. Idiomatic expressions are combinations of words which have a peculiar grammatical function and whose components do not give a clue to the meaning of the whole. For example, "in order to" as a unit can be called an infinitive particle and although its meaning is clear to us, the words taken separately do not contribute to an understanding of the expression. Colloquial expression is the conversational way of saying things which may or may not violate formal rules of grammar. Both idiomatic and colloquial expressions present great problems for machine translation because they do not follow defined linguistic patterns and move away from the norms of written language. Technically, colloquialisms should not be a problem because the nature of the texts targeted for machine translation would seem to preclude the use of conversational language. This, however, is not really the case. Non-conventional spoken language finds its way into even the most highly technical literature. We must keep in mind that the technical writer is not mentally distinguishing between formal written and spoken language, but rather is trying to relate instructional messages and will use whatever language seems familiar to him or her. Idiomatic expressions are definable to a certain extent. Pre-edit draws the most common to our attention immediately, and these can be assigned parts-of-speech and be entered in our translation dictionaries as units. However, we can never saturate our dictionaries with all possible idiomatic expressions. Another characteristic of today's technical literature, especially computer technology, is the use of "computereze" which is so highly idiomatic and always changing that it defies any attempt at formalization, presenting another stumbling block for machine translation. Idiomatic and colloquial expressions have a place high in the hierarchy of difficulty, because they reflect social patterns of communication which are subordinate to human requirements. This way of expression is not necessarily a result of thought processes, but more a reflection of style and the social state which is constantly in a state of flux. Who can really say that the expression "to the max" will not find its way into written language and even into technical literature? And no doubt the future holds even more linguistic surprises. We should also consider the fact that even if we could devise a system "to harness" current idiomatic and colloquial expressions, providing equivalents is a tricky task, as it would require a knowledge of current expressions in the target language to be able to give an equivalent that truly reflects the expression in the source language.

Style and contextual considerations

This last level on our pyramid is the most complex and the most elusive. Control of style to fit writing conventions that lend themselves to accurate analysis by the computer is a very difficult task for the person working with machine translation. When we speak of saturation in relationship to style, we are referring not to dictionary saturation, but rather system program saturation. The question here is whether or not it is possible to achieve such a thorough set of source language analysis and target language synthesis programs as to be able to cover all possibilities of human expression. The answer of course is negative.

In addition, control of input, or rather the restriction of style, is very difficult to accomplish. There are certain stylistic aspects that we can attempt to control. Typically, passive voice is discouraged when English is the source language, not so much because of problems it presents to analysis, but because of the problems it causes in target language synthesis. Stylistic variances often cause more problems for target language composition than for source language analysis. In other words, when it comes to questions of stylistic problems the computer often has no problem in parsing the sentence. The real issue becomes putting that sentence into a correct and acceptable style in the foreign language. It is really a challenge to try to teach someone to alter style of writing for the computer. It is very difficult to teach a writer to use his or her language in a restricted fashion. Those of us who work closely with machine translation can easily spot a style or structure that is going to be a problem for the computer, but it is very difficult to define these and then train the writer to avoid such "ways of saying things" in their texts. Few controls can be put on style because so much is acceptable in the English language. Style is also determined by affective qualities that cannot be defined and therefore controlled. Style is determined by and meaning is assigned through context. As the machine has no contextual frame of reference, to translate the style and meaning into another language becomes impossible. To the extent that we can control the input and correspondingly saturate our system with linguistic programming to give us an acceptable output, we can claim to have some success, however limited, on this highest level of our pyramid.

Conclusion

Translation by machine can produce good results if those who work in this field realize that control and saturation can compensate to an extent for the intricacies of language and the enigma of human thought processes. The hierarchy of difficulty of control that is presented here is not intended as a definitive statement on the subject of controlled language for machine translation, but rather is intended to point up some of the difficulties we encounter when we attempt to build a translation system and to put controls on natural language. We will realize the greatest benefits from machine translation when we admit to its limitations, and we will have the possibility of overcoming some of these limitations when we begin to focus our attention on them.