4-6-1979

Its - An Interactive Translation System

Alan K. Melby

ITS - AN INTERACTIVE TRANSLATION SYSTEM

Alan K. Melby

It is well known that computers are tremendously useful in solving long and involved arithmetic calculations. We are now seeing an explosion of additional application areas for computers. The Translation Sciences Institute (TSI) of BYU is exploring various ways that computers can be made useful in translation.

FAHQT - A CURRENT IMPOSSIBILITY

The idea of machine translation (MT) is not new. There have been many MT projects since the 1950's. Nearly thirty years of research on this problem have pointed out the immense difficulty of the task, the inadequacies of linguistic theories, and the limitations of computers. It is now generally agreed that fully automatic high quality translation (FAHQT) of general text is impossible, given our current, limited understanding of language. However, this does not mean that computers are not extremely useful in translation. It simply means that the computer is better suited as a partner with, rather than a replacement for, the human translator.

REACTIONS

The realization that FAHQT was not soon to be obtained evoked widely varied reactions.

1. Basic Research

Some have turned their attention to basic research. For example, some workers in Artificial Intelligence do automatic translation restricted to a small set of structures and vocabulary items. They have no immediate plans for large scale application, but they are doing very interesting work in the context of "a general investigation of language and thinking" (Carbonell et al, 1978, p. 50).

2. Machine-Aided Human Translation

Others are interested in commercially practical machine aids to translation. One option is to develop special editors and automated dictionaries for translators. These aids have already proven their usefulness and are becoming more widely accepted and used (at TSI and elsewhere). This level of machine assistance can be termed machine-aided human translation. At this level the human takes the active role and requests specific help from the machine. There is another level of man-machine partnership which we will call human-aided machine translation. At this level the machine takes the active role and requests specific help from the human. This human help can take several forms.

3. Human-Aided Machine Translation--Montreal

One MT project (METEO) at the University of Montreal translates weather forecasts from English to French. The MT system examines each sentence individually. It translates some sentences on its own. When it runs into trouble on a sentence it sends it to a human translator at a video terminal. The system then merges the machine-translated and human-translated sentences into a single document which can be reviewed and distributed. This approach has been successful in the translation of weather reports because they form a rather restricted sub-language (Kittredge, 1978). This allows the programs to be tailored to the sub-language in question, and 60% to 90% of the sentences of weather report can be translated automatically. The sub-language approach is currently being applied to the translation of a certain kind of aviation manuals.

4. Another Form of Human-Aided Machine Translation--BYU

The BYU MT project (ITS) determined over seven years ago to develop a human-aided machine translation system. However, a different set of requirements led to a somewhat different implementation than the Montreal project. The BYU project was to handle general modern English prose, especially LDS Church publications. This material is definitely not restricted enough to be called a sub-language. An automatic translation would be acceptable on only a fraction of the sentences of a typical text from general or LDS English. So the BYU group decided to have the computer ask the human for help on specific problems within each sentence during the course of the translation process, rather than requiring each sentence to go fully automatic or fully human. This has led to years of interesting research on the problem of how to get the computer to know what to ask the human and what to do with the answer. The unifying linguistic model of the BYU project is Junction Grammar.

There is another important requirement on the BYU system. It must produce output in multiple target languages. A reasonable question to ask is whether the computer need ask a separate set of questions as it translates into each language or whether the computer can ask questions which are helpful in translating into several or all of the target languages. This sharing of questions does in fact work and allows the overhead of human interaction to be distributed over all the target languages. Thus, for each added target language the number of additional questions becomes smaller and the system as a whole becomes more cost effective.

5. All or Nothing

We have seen that when it was realized that fully automatic translation would not come soon if ever, some returned to basic research, some developed machine aids for human translators and some explored human-aided machine translation. Unfortunately, there were also some who concluded that if computers could not do fully automatic translation of all kinds of text, they were of no use at all in the translation process. This is like saying that if an automobile cannot drive by itself over any terrain it is not useful at all. Those who insist on FAHQT should by analogy insist on being able to tell their car their destination and sit back.

They should reject automobiles entirely because of the need of human aid at the steering wheel.

The BYU group feels that even with its limitations and even if it is only in the Model-T stages, human-aided machine translation can and will be very useful in solving real world communication problems and will produce many significant insights into the nature of language as well.

The rest of this paper will describe the BYU Interactive Translation System (ITS) in general and a few of the recent advances in the system.

BYU ITS SYSTEM

In a production environment, the input to the BYU system will normally be a document which has been or is being published in English and is to be translated into all or several of the target languages of the system. Currently ITS translates from English into Spanish, Portuguese, German, French, and Chinese. At present, the major effort is adding more grammar and vocabulary, including idioms. The current version of the system was begun a little over a year ago and is scheduled to be ready for production testing in September 1979. Until then, most test material is oriented to the specific grammar and vocabulary items being programmed. In addition to our full-time activities of adding grammar and vocabulary, the system is already used to translate a paragraph a week of current LDS English into the five target languages to monitor the system and detect problems. Of course, the paragraph is pre-edited to remove constructions which have not yet been programmed.

The major steps in the ITS process are: SETUP, ANALYSIS, TRANSFER, SYNTHESIS, and POST-EDITING.

1. SETUP

SETUP is the process of defining manageable blocks of text and numbering the sentences within each block. In the case of current LDS publications, the English text is available on a typesetting tape. To avoid unnecessary retyping of text, we have recently developed the capability of automatically decoding the typesetting tape into format which can be fed directly into the ITS.

2. ANALYSIS

After SETUP, the text is analyzed, sentence by sentence, into a representation called a J-tree (i.e. a Junction Grammar tree). The J-tree makes explicit many aspects of language that humans determine unconsciously when they listen or read. For example, modifiers point explicitly to whatever they modify and ambiguous words receive special suffixes to indicate which definition applies in the sentence in which they occur. J-trees are defined by Junction Grammar, which was developed by Eldon Lytle.

During analysis, the computer must often ask the human operator for assistance. This is done using a video terminal. However, as mentioned

before, the questions of analysis are "shared" questions. Thus, analysis is only done once, no matter how many target languages the system will translate into.

3. TRANSFER

The J-tree produced by ANALYSIS has neutralized many of the apparent differences between languages. Therefore, given a sentence and its translation into, say, German, suppose one analyzes both the original sentence and its translation into J-trees. The differences between the two J-trees can be mechanically adjusted far more easily than the differences between the surface sentence and its translation. The task of TRANSFER is to adjust for the differences which remain at the J-tree level.

Most of the adjustments of TRANSFER can be done automatically. They are stimulated by the presence in a sentence of a particular word or structure. But another task of transfer is doing the actual translation of words from English into a target language. In many cases, even where the words are ambiguous in the original sentence, the choice of a translation can be made automatically. However, there are cases of "precision resolution" where the information in the J-tree is not precise enough to determine the proper translation of a word. In the past, these cases were handled by requiring TRANSFER to make an arbitrary choice. If the choice was wrong, it was cleaned up by the post-editor. However, it was observed that this approach often required more adjustment by the post-editor than just changing the base form of the word. No word is an "island", so to speak, in a sentence. The choice of one word often affects the inflection and/or order of one or more other words in the sentence. So the effect of choosing the wrong word in TRANSFER is often magnified several times by the time it reaches the post-editor. This is the "precision" problem.

Recently, we realized that the post-editor may use his time more effectively if the computer asks some precision questions in TRANSFER before the effect of a wrong choice is magnified. The computer narrows down the possible translations of a troublesome word as far as possible and then presents the remaining choices to the human operator. The human need only reply with a single number to indicate his choice rather than changing a whole word and its implications later on. Further experimentation and "tuning" of the system will be necessary before an optimum balance of human-machine interactions in the various steps will be obtained.

4. SYNTHESIS

SYNTHESIS is the processing step which converts an adjusted J-tree into a sentence of the target language. It is very important but often goes somewhat unnoticed because its processing is entirely automatic.

5. POST-EDITING

The POST-EDITING step is a review step where the completed translation is examined by a human post-editor (probably the same person who

answered the precision questions in TRANSFER). After post-editing, which could include a second review at a translation center in the target language area, the translation can be typeset automatically. It is extremely significant that during the entire ITS translation process the text need never be manually entered or retyped. During POST-EDITING, only the parts which need changing are changed. This eliminates many typographical errors.

RECENT ADVANCES

There are many recent advances in the ITS which were not discussed above but deserve mention. I have chosen only five.

1. Computer Resources

It has been almost exactly once year since the dedication of the Wilkinson computer, an IBM 370/128. The machine has performed exceptionally well the past year and has greatly enhanced the institute's computer resources. It has allowed us to convert to the VM operating system, which has facilitated program development. The Wilkinson computer has also been tied to the institute's NOVA 3 mini-computer by direct wire, thus further integrating the ITS and word processing aspects of the institute. Our profound appreciation toward the Wilkinson family continues.

2. Scripture References

When scriptures are quoted in LDS publications, they are not retranslated with the rest of the text. The official Church translation is looked up and inserted. This is a task which takes no imagination and should be done with a computer. We wanted to do something but did not have the resources to enter the entire set of Standard Works in five languages immediately, so we hit upon a compromise plan in cooperation with the BYU library. The library has kept careful records of all the scriptures quoted in General Conference since 1950. Using TSI programs and with the help of library personnel for data entry, the library's entire scripture reference file has been entered onto disk files and processed to produce various listings. The ITS language teams are now entering the official translations of all the scriptures which were referenced five times or more. This amounts to about 1300 verses. These verses account for about 60% of all the scripture citations in General Conference in nearly 30 years, yet they amount to only 2% of the verses in the Standard Works.

3. Corpus and Concordance

There are many projects which have produced a corpus of some sort and a concordance of it. The concordance is a well-accepted tool in literary analysis and gospel study. Upon consultation with the Montreal translation project, which uses concordances for linguistic research, we decided it would be worthwhile to gather a specialized corpus of modern LDS English. We chose material from the Church magazines, Sunday School manuals, Family Home Evening manuals, etc. Using the decoding process mentioned above, we were able to produce a corpus using typesetting tapes borrowed

from the Church word processing service in Salt Lake City. We then developed a technique for producing a file of pointers from each distinct word to its occurences in the corpus. To this we added a means of dynamically generating a subset of the concordance for a specific list of words that a particular researcher is interested in. Note that all this was done without entering the data manually.

The corpus currently consists of over one half million words and can be enlarged as needs and resources dictate. It has been available to institute personnel for only a month but is already being used regularly in researching questions of word sense, idiom, and function word usage.

4. Special Characters

Thanks to the integrations of ITS and the word processing activities at the institute, we can now obtain proof copies of translated output in "true" characters (i.e. upper and lower case characters with diacriticals) on a special print train at several hundred lines a minute. We can also obtain proofs on a "daisy wheel" printer for slower but higher quality output. Importantly, the same disk file which is read to produce proofs can then be read to produce typeset output if the proofs are found to be satisfactory. Due to other recent advances, we can even obtain proof copies of output for languages that do not use the Roman alphabet, even for Chinese and other non-alphabetic languages. During the regular ITS processing, the Chinese characters are represented by telegraphic codes as specified by CETA, a Washington D.C. organization which correlates such matters in the United States. Then the telegraphic codes are sent to the NOVA 3, converted to a dense matrix of dots, and printed on a Versatec printer/plotter.

5. Linguistics and Interaction

There has also been progress in the area of linguistics and the interaction between the human operator and the computer. The first priority in ITS is to get enough grammar and vocabulary into the system so that it can handle real text with little or no pre-editing. That means the system must produce a detailed syntactic-semantic representation of each sentence. This has not yet been done but, if all goes well, will be nearly attained by this fall, thanks to advances in the range of constructions that are now handled by Junction Grammar. Only a few years ago, a full analysis of real text could not even be done on paper. We would pick up a Church magazine and start analyzing a paragraph. Nearly half the sentences would contain constructions that were still not treated satisfactorally in Junction Grammar. We could only hope that further linguistic research would find the needed answers. That hope has largely come true. Now it remains to program the grammar we understand.

Of course, once the system is able to handle real text, it is important to reduce the interaction needed during the translation process where possible without sacrificing the quality of the output. Until recently, the cost of the computer processing was a major concern. But as the cost of computer time continues to drop and the cost of human translator time increases, it becomes desirable to have the computer go

to more and more trouble to answer some of its own questions, and to make the questions it does ask easier and faster for the human to answer. We have recently taken some steps in this direction.

a. Dictionaries

In the area of dictionaries, the various kinds of information they supply is being redistributed to make the computer processing more effective and to make the interactions easier and more accurate.

b. Guesses, Confidence, and Interaction Level

The analysis and transfer programmers are no longer forced to decide between interacting always or never on a given question. We have added the option of interacting "sometimes," depending on a delicate dance of the computer's "guess" at the answer to its own question, its "confidence" in its guess, and the "level" of interaction specified by the operator. At a lower interaction level, the computer will ask fewer questions but make more mistakes, and vice versa at a high interaction level. This capability will allow us to experiment to find optimum settings for the various parameters. The effort is somewhat like optimizing the performance of a complex machine.

c. Ease of Answering Questions

In terms of overall human time involved, it is just as effective to make it easier and faster to answer questions as it is to reduce the number of questions. We have just received new video screens which display more lines of text and have developed high-lighting techniques which should allow progress in this direction.

6. Literal Translation

As mentioned above, some sentences of real text contain constructions which have not been programmed for. Due to the tremendous flexibility of human languages, this may decrease but will never go away. Therefore, we are implementing a "literal" option in this version which steps in to avoid total failure on a sentence and provides a simple word-for-word translation for the post-editor.

FUTURE PLANS

There are several areas where much more research and development on the ITS system will be needed even after this fall. I will mention only a few. In addition to improving the grammar of English implemented in analysis, it is anticipated that there may be added the capability of looking further ahead and pursuing the consequences of several possibilities simultaneously, thereby reducing interaction. This capability is heavily used in some MT projects, while others claim it is needed only on

a limited basis if all aspects of a semantic analysis are properly integrated. A more obvious need is for the system to analyze text as coherent discourse rather than as isolated sentences. This would allow the system to take advantage of answers the human provides early in the text to make better guesses later in the text. Discourse analysis would naturally lead into the problem of drawing inferences from statements in the text and eventually would lead to the problem of somehow representing the real world inside the computer in a useable way. All this has been done in a very restricted context called a "micro-world" but never in a large-scale system such as ITS. Of course, as the analysis grammar is improving, transfer and synthesis will need much more research and development.

While further research is in progress, the institute finds itself in the desirable position of having found several levels of useful machine-aids to translation without having nearly exhausted the possibilities for major future advances.

REFERENCES

Carbonell et al., 1978, Proceedings of the 1978 International Conference on Computational Linguistics, held in Bergen, Norway.

Kittredge, 1978, Same as above.

To obtain more information on Junction Grammar and ITS, write for a copy of the journal Junction Theory and Application. Selected issues contain an annotated bibliography.

Junction Theory and Application
Attn: Subscription Editor, Jill Peterson
Room 130 Building B-34
Brigham Young University
Provo, UT 84602