Computer analysis of basic English as a first step in machine translation

J.R. STRAUB† and C.A. ROGERS*

† University *of Arizona, Tucson, USA*
** Augusta College; Georgia, USA*

The results which have been obtained from machine translation have not yet approached the promise which the early researchers saw as being easily obtainable (Booth [1]). The speed of the computer,, as well as its ability to store and retrieve words in a dictionary, were thought to be sufficient to develop mechanical translation. Problems of grammar and syntax could be handled either by clever programming or by some combination of entries in the computer's dictionary.

The researchers who thought that they had sufficient resources available in the computer did not reckon with the difficulties of translating a natural language by machine (National Research Council [2]). Obstacles soon became apparent: irregularities and inconsistencies in word order between the source and the target language; idioms and, above all, ambiguities arising both from grammar and from semantics all seemed to have no neat solution. Despite recent improvements in the technology of data processing, such as increased processing speed and increased on-line storage capacity, in practical terms three serious deficiencies have continued to prevent machine translation from becoming a useful by-product of the computer. The quality of the translation has remained poor, the speed of translation has remained low and the cost has remained high.

With these considerations in mind, we attempted a different approach to machine translation. Researchers had, in the past, made use of human pre-editors who scanned the computer's input and ' added data to eliminate or reduce problems of word order, idioms and ambiguity (Oettinger [3]). What we proposed to do, in effect, was to transfer the function of pre-editing to the speaker of the source language. If a language could be found which a computer could analyze without errors in word order, idiom and ambiguity, it would seem that some obstacles to machine translation would be eliminated.

The language we chose for this study is a subset of English, called Basic English. Although restricted in vocabulary and grammatical rules, it is a natural language which has been learned by many speakers and is used in normal human communication (Richards [4]). Basic English is the formulation primarily of Charles K. Ogden and was developed during the 1920's. It employs a vocabulary of 850 words, divided loosely into 100 operations, 600 things and 150 qualities. Only 18 of the operation words arc English verbs and they convey motion, possession or attitude. The rules as verbs in Basic English. A word which is a thing in Basic English, for example, 'answer'...may never be used as a verb.

With each of its words clearly defined in terms of type and function, Basic English seemed to be well-suited to computer analysis. The problem which has plagued researchers in the past, ambiguity, should decrease within the rules of Basic English. We postulated that a computer program could be written which would correctly analyze all the elements within a Basic English sentence, with no ambiguity. If we demonstrated that such a program could be constructed, we would also demonstrate that Basic English was capable of being translated by machine with little or no error arising from the source language. Complete machine translation would require the development of synthesis programs to generate sentences in the target languages. In order to have the same non-ambiguous conditions in the machine output, it will be necessary to make use of 'basic' versions of the target languages. Such a complete system awaits development; this project was aimed at demonstrating the exactness with which Basic English may be analyzed.

The computer system

Hardware/Software

The hardware which was used to develop and implement the system was a GE-415 processor with 32K words of memory. There was no on-line storage available on this system; the only high-speed storage devices available were 5 magnetic tape handlers. The system also contained a card reader, a card punch and a high-speed line printer.

The programs were written in COBOL for use under the GE Basic Operating System, a stand-alone, batch environment. Every effort was made to utilize COBOL in such a way that conversion to the COBOL of other manufacturers could be done with a minimum of reprogramming.

The basic English analyzer system design

We saw that there were two separate tasks which the Basic English Analyzer System (shortened to BEAST) would perform: first, it would build and maintain a dictionary of Basic English; and, second, it would accept and analyze Basic English statements. Based on this functional analysis, BEAST was designed as two separate programs, a dictionary program and an analyzer program. The dictionary program is used to create a valid dictionary of Basic English words, and can produce listings of the dictionary on demand. The analyzer program uses a. magnetic tape of the Basic English dictionary to produce a printed analysis of Basic English state ments which are submitted on punched cards. BEAST is designed so that it can be used by a speaker who is not yet an expert in the rules and vocabulary of Basic English. The system recovers fully from any errors which may be present in the input statements by noting the error and returning to the starting point to process more statements. This facility will be especially pertinent if BEAST is modified to run in real-time with conversational . input.

The present version of BEAST has the following capabilities:

1. BEAST will recognize any word in the Basic English vocabulary, with no ambiguity.
2. BEAST will recognize and process irregular forms of words, such as past tenses and participles.
3. BEAST will recognize and process variants of words, such as irregular plurals ('foot'...'feet') and letters which are lost by English spelling conventions ('move'...'moving').
4. BEAST will determine the number of nouns.
5. BEAST will determine the tense and number of verbs, including irregular forms.
6. BEAST will determine the degree of adjectives.
7. BEAST will determine whether an adjective is used as a noun.
8. BEAST will determine whether a statement is indicative, imperative or interrogative.
9. BEAST will determine whether a statement is a complete sentence; that is, whether it con tains a valid verb.
10. BEAST will differentiate words by their usage; that is, 'to' in ah infinitive construction is differentiated from 'to' in a prepositional phrase.

The restrictions which are present in the current version of BEAST were engendered by the limited hardware, which was available for the system's development. These restrictions arose from the small amount of core memory and the unsophisticated peripherals of the GE-415: .

1. BEAST cannot currently analyze input state ments unless they are punched on cards in a rigid format.
2. BEAST cannot accept a single statement of more than 30 words.
3. BEAST cannot now develop relationships from one statement to another.
4. BEAST cannot now be used in a time sharing environment, nor can it be operated in a teleprocessing environment with data entered at a

remote station.

Program descriptions

The hardware requirements for the dictionary program are a card reader, a line printer and, depending on the selected input/output option, 1 or 2 magnetic tape handlers. The dictionary program requires as input a control card and dictionary data, either card or tape. The dictionary program produces as output a magnetic tape and a printed listing of the Basic English dictionary. Depending on the options indicated on the control card, the program will read the dictionary data either from the card reader or from a magnetic tape. The data is edited and codes which indicate the Basic English part of speech are verified. A printed listing is produced and the dictionary is written onto a magnetic tape. It is this output tape of the dictionary program which is sorted alphabetically and used by the analyzer program in the look-up process.

The hardware which is required by the analyzer program includes a card reader, a line printer and 5 magnetic tape handlers. Input to the analyzer program consists of Basic English statements, which are read from the card reader and the Basic English dictionary, which is located on magnetic tape and consulted for each statement read by the program. Output is a printed listing of the input statements as they were read by the program, a listing of the intermediate results of the program's analysis and a listing of the Basic English statement.

The analysis program consists of four main phases which are executed one or more times until the analysis is complete. In Phase 1 (Input Phase) the program reads and stores in memory the Basic English statement which is to be analyzed. This phase edits the input statements, truncates oversize statements and lists on the printer all the data which it has read and stored. Phase 2 (Sort Phase), notes the position of each word within the input statements, sorts the statement in memory into ascending alphabetical order and prints a diagnostic message when it is complete. In Phase 3 (Analysis/ Match Phase) the program uses the alphabetical grouping of words in memory to perform an item and syntactic analysis of the input statement. The order in which this is accomplished is the following:

1. A preliminary scan for endings and affixes, which are removed so that the word's stem can be located in the dictionary.
2. A matching process involving the ordered words in memory, the work tapes and the dictionary tape.
3. An additional pass against the dictionary tape to resolve irregular forms and spellings.
4. A final analysis to interpret syntax in view of the dictionary information which has been obtained in the matching phase.

In Phase 4 (Output Phase), the program reconstructs the original statement, using the analysis data which has been determined in the earlier phases. When this loading process is complete, the program prints the final statement and the analysis data on the output listing. Words which have not been processed are deleted from the final output but are annotated with the message 'item dropped'. The program identifies the statement as being imperative, indicative or interrogative, and whether it contains a complete verb form.

When the analysis is complete, a message is printed, the program returns to its original internal state and the analysis of the next statement begins.

The system as it is currently configured is mainly constrained by the speed of the magnetic tape • handlers. Statements ranged from one word to 30, and were processed in from 10 seconds to 35 seconds each, with an average processing time of 20 seconds per statement.

Expansion of BEAST

BEAST can be modified to run even more quickly, to process larger statements and to develop relationships between statements if the system's hardware is altered. Increased core memory and replacement or supplementation of magnetic tapes with disk or other direct access storage media will provide the hardware framework within which these software revisions can be accomplished.

In terms of BEAST purely as a computer system, there are several enhancements which could easily be implemented in an improved computer environment. Data could be entered into the system in a less rigid form than the current format rules allow. Punctuation marks, aside from the terminator, should be allowed within the body of the text. Literal data should be allowed to pass into the system without being located in the Basic English dictionary. In the present system, for example, names of people and places are dropped because they are not found in the Basic English vocabulary.

The translation system is planned to operate in an on-line interactive mode. This goal can be reached only after the hardware has been enhanced to allow a teleprocessing system to be installed. BEAST can then be modified to input and analyze statements from a remote typewriter terminal or CRT. Future synthesis programs will also be written as teleprocessing systems, eventually allowing conversations, through the computer, from Basic versions of one language into Basic versions of others.

Conclusions

Even with the restrictions and shortcomings which the current version of this system possesses, BEAST has demonstrated that a subset of a natural language exists which can be analyzed successfully by a series of computer programs without producing ambiguous results. The ability to produce non-ambiguous results lies not in any special programming algorithm or method of constructing a computer dictionary, but instead is dependent on the non-ambiguous nature of Basic English. It is from this quality of Basic English that BEAST's results are obtained. The function of eliminating ambiguity has been shifted from the position of a pre-editor, working at a point between the English speaker and the computer, to the English speaker himself. All that the speaker need do is to speak, or write, within the rules and vocabulary of Basic English. If the speaker uses Basic English correctly, the language itself will automatically eliminate enough ambiguity so that the speaker's statements can be correctly, quickly and cheaply analyzed by the computer.

In order for the translation process to be completed by computer, a 'basic' version of the target language needs to be reduced to the form of computer programming algorithms and a machine readable dictionary. A hypothetical system such as this would have as its input the analyzed results which BEAST has generated from a Basic English input statement. The synthesis process would be the reverse of analysis, with output statements built up from the rules and dictionary of the 'basic' version of the target language. Such a system awaits experimental verification.

With expanded hardware resources, both analysis and synthesis programs will be teleprocessing in nature. Such a translation system will allow.for live, on-line translated conversations among two or more terminal users, who may be at any distance from each other, as long as they are in telephone contact. Remote communication systems are readily available today; only a language synthesizer remains to be developed before remote language translation may be attempted.

Earlier researchers had suggested that an 'intermediate language', simplified and trimmed of vocabulary, needed to be developed in order to reduce computer error (Panov [5]). The use of Basic English in this capacity shifts to the speaker the task of choosing non-ambiguous words from the basic vocabulary. The speaker's own statement is equivalent to this non-ambiguous intermediate language, which is ready to be translated into one or more output languages.

The development of BEAST has shown a practical method of circumventing the ambiguity errors which have been a shortcoming of machine translation. Since ambiguity arising from the source language has been a barrier to machine translation, the development which this system represents can be considered to be the beginning of a fruitful machine translation process.

References

1. BOOTH, A.D., *Aspects of Translation,* Secker and Warburg, London, pp.88-92 (1958).
2. National Research Council, Automatic Language Processing Advisory Committee, *Language and Machines,* National Academy of Sciences, Washington, D.C., pp.16-18 (1966).
3. OETTINGER, A.G., *Automatic Language Translation,* Harvard University; Press, Cambridge, pp. 114-118(1960).
4. RICHARDS, I.A., *Basic English and Its Uses,* Norton & Company, New York, pp.21-44 (1943).
5. .PANOV, D., *Automatic Translation,* Pergamon Press, New York, pp.62-64 (1960).