[From: *Information retrieval and machine translation*, ed. Allen Kent. International conference… September 6-12, 1959, Cleveland, Ohio; part 2 (New York: Interscience Publishers, 1961)]

CHAPTER 27

Machine Translation at the University of Michigan*

ANDREAS KOUTSOUDAS

Machine Translation Project, The University of Michigan, Ann Arbor, Michigan

I. INTRODUCTION

This paper is intended to give a general idea of the methodology and aims of the research in machine translation (MT) being carried on at the University of Michigan. Before I begin, however, I think it worthwhile, in view of the nature of this conference, to make a few remarks on the question of hardware for machine translation.

The main fact to be borne in mind, it seems to me, is that, because of the infancy of machine translation, it is very hard to say how well existing machines measure up to all its eventual needs, or what modifications would have to be incorporated in a machine designed specifically for translation. It is true that considerable weight is sometimes given to the fact that machine translation will require a computer with very large immediate access storage, but concern with this sort of problem seems premature. After all, storage media of indefinite capacity, such as magnetic tape, do exist, and it is an open question whether or not the slow access time of such media would create difficulties for translation; e.g. a bottleneck at the stage of dictionary lookup.

The main task of machine translation research, at present and in the foreseeable future, is the linguistic task. Until we know in detail how to solve the problems of lexical ambiguity, inflection, and syntax (discussed below), for a given pair of languages, we shall be unable to construct a complete computer program for translation. And until such a program can be written, we can only guess at the problems which would be involved in adapting a computer to handle it. At present we are working at Michigan towards a program capable of translating Russian texts in the field of physics into English. We are pursuing two separate approaches, which may or may not turn

* The present work is being conducted at the University of Michigan with research funds provided by Project Michigan, under U. S. Army Signal Corps Prime Contract Number DA-36-039— SC-57654.

out to be interrelated, but which both offer interesting avenues of search. These approaches can be roughly labeled "empirical" "formal." I shall describe them in that order.

II. THE EMPIRICAL APPROACH

The main linguistic problems facing machine translation are those of lexical ambiguity, inflection, and what might be called "nonequivalence of syntax." Lexical ambiguity arises from the fact that a given word in one language may have more than one possible translation in another language. When a lexically ambiguous word occurs in a sentence being translated, some provision must be made for the machine to select, from among the alternative translations open to it, the one which is required by the context of the given sentence.

An inflected form consists of a stem to which endings are attached to convey grammatical information. For example, in played, play is the stem, and -ed, conveys the past tense. In English, inflection is comparatively unimportant, but in a language such as Russian inflection assumes a much greater importance in conveying grammatical information and therefore it can present considerable problems in translation. For example, the machine must be capable of determining the stem of an inflected form, of abstracting and utilizing the grammatical information conveyed by the ending, and of handling cases in which the grammatical function of a word is not clear from its inflectional ending alone, since in some cases endings overlap This last problem is essentially a variant of the problem of lexical ambiguity.

Generally speaking, syntax is the process by which words combine to form phrases and sentences (for example, the distinction between man bites dog and dog bites man is a syntactic one). Syntax is troublesome because different languages use different word order patterns to indicate the same grammatical relationships. A translation in which the separate words have been translated but the original word-order retained, may be unintelligible, or worse, plausible but incorrect.

One way of tackling all these problems is to try to discover rules based on diagnostic "clues," or "signals," in the written context surrounding problematical words or constructions, which the machine can use to reach a decision. For example, the English translation of the Russian word zavisimost' may be either "function" or "dependence." If we find out, for example, that in every case, when zavisimost' is followed immediately by the preposition ot, the correct translation of zavisimost' is "dependence" then we may be able to use this fact, with others, to formulate an instruction which will enable the machine, in a high percentage of instances, to distinguish the correct translation of zavisimost'. There are several ways in which one might go about developing rules of this kind. For example, we might, as Y. Bar-Hillel (1) has recently advocated, try to derive

them from already existing descriptions of languages. Bar-Hillel argues that in view of the considerable developments in descriptive linguistics over the last fifty years, it is absurd for MT workers to feel obligated to begin studying language entirely afresh for the purposes of translation, particularly in the case of languages (such as Russian) which have been fairly exhaustively described.

There are several reasons, however, for treating Bar-Hillel's whole contention with caution. To begin with, it may be questioned whether the results of descriptive analysis are completely adaptable to the requirements of computers. It is the ambition of at least some structural linguists, to avoid semantic considerations and to be as "formal" as possible, and no doubt this ideal is to some extent achieved. But the point must be made that only insofar as it is achieved can structural linguistics be relevant to machine translation. Furthermore the aims of the descriptive linguist and the MT worker are rather different. The linguist wishes to elaborate, or discover the simplest possible set of procedures (rules) which will generate, or describe, all and only, the constructions of a given language. The worker in MT on the other hand, is concerned with rules for generating, with respect to any given sentence of a language, its equivalent in another language. In other words the linguist analyses single languages; the MT worker pairs of languages. It might still be argued that, given complete descriptions of the syntactic and inflectional structures of two separate languages, it would be a comparatively simple matter to devise rules connecting one set of structures with the other. But now, first of all, even granting the existence of structural descriptions of the requisite degree of completeness, the work of devising rules to "connect" them would still have to be done by the MT worker (i. e. descriptive linguists would be an aid to the development of translation rules, and not itself a source of such rules). And secondly, it is my opinion that the existing descriptions of languages are not nearly so complete as Bar-Hillel suggests, particularly where syntax is concerned. Even for a language such as English a considerable amount of disagreement exists as to what constitutes a correct analysis.

Finally, descriptive analysis would be relevant only to the problems of syntax and inflection, and would hardly help us to find rules for distinguising, on the basis of context, between the alternative meanings of lexically ambiguous words.

A second method of developing rules might be for a person who knows two languages very well to derive a set of rules for translation on the basis of his own "linguistic intuition," and proceed to test them against samples of the sort of text which is desired to translate.

The difficulty with this is that to formulate a set of rules sufficiently comprehensive to stand up under testing at all, one would have to be only fluently bilingual, but trained both in linguistics and in the subject-matter (e. g. physics, mathematics) of the material to be translated. And these qualifications are rarely met within the same person.

The approach that we have adopted differs from the above methods, Its main feature is that in formulating a preliminary set of rules to handle a particular problem (the set, that is, which must form the basis of all further testing and modification) we rely, not on linguistic intuition* or pre-existing descriptive analysis, but on the study of very large texts (of the order, at present, of 73,364 running words). This has the immediate advantage that one can be sure, even before testing, that one has a set of rules of fairly wide applicability. Again, the relative frequency of "exceptions" can be immediately evaluated.

The process of deriving rules from a text can be best described by an example. Consider the problem of lexical ambiguity. It seems clear that the only way to enable a machine to select a correct translation for any given word from a group of possible alternative translations, is to enable it to refer to the immediate context of that word for diagnostic clues or "signals," (words, punctuation marks, etc.) There are, moreover, reasons for thinking that, for most cases of ambiguity, clues are to be found within a context of not more than two or three words on either side of the ambiguous word (2). It seems reasonable, therefore, to begin by determining how far ambiguity is resolved by a context of one word on either side of the ambiguous word, extending this in further experiments until all ambiguity is resolved. The first step in this direction is to abstract all the ambiguous words from a very large sample of scientific text, together with the two-word context of the word on either side of the ambiguous word for each word. Given this material we can design an experiment which will reveal diagnostic clues for a proportion of ambiguous words. Later on, rules can be formulated in terms of these empirically discovered clues. Because of the size of the text from which they have been derived, the inadequacies of these preliminary rules are fairly apparent, and further experiments can be planned in the light of this information. The general pattern of all our empirical studies is roughly the same as that of the above example, in that rules are formulated in the light of accumulations of relevant information extracted from very large texts. More precise information about the results which we have achieved so far will be found in the Bibliography.

III. THE FORMAL APPROACH

We now come to the second, or "formal" approach, which, essentially, is concerned with the development of "learning" or “self-modifying" programs for machine translation. So far we have discussed methods for developing what I shall call a fixed program. The rules comprising such a program would be derived from, and therefore adequate to translate, a certain finite text, and assuming the

*This is not to say that we make no use at all of intuition. The advice of native speakers is used to resolve specific questions, such as the meaning of words.

this text was large enough, would also do a more or less adequate job of translating further texts. But it is clear that however large a text we base such a program on, occasions will arise, as we apply it to more and more new texts, when it will cease to be adequate, and will therefore need to be modified.

Now, there are two obvious ways of modifying a fixed program. One might allow inadequacies of translation to pile up, and reprogram the machine completely at intervals. This would be clumsy and time-wasting, since at each reprogramming the complexities would probably increase. On the other hand one might try to develop a "master" program capable of modifying the translating ("slave") program whenever an inadequacy arises. Such a system would have to meet the following requirements:

(1) Suppose the slave program to be such that it translates the series of sentences $S_1 \ldots .. , S_{n-1}$ correctly, but the sentence S_n incorrectly. Now, it should be possible for the master program, given the input sentence S_n and its (corrected) translation S'_n, to modify the slave program in such a way that in the future, when presented with S_n, it produces S'n instead of the incorrect version.

(2) However, the modified slave program must continue to translate $S_1, \ldots, S_{n-1}$ correctly, or nearly so (i. e. modifications must be so carried out, that a limited series of modifications will result in the machine translating all of $S_1 \ldots\ldots S_n$ correctly).

(3) Modifcations should not lead to the program eventually becoming over-complex and clumsy (e. g. modifications should not take the form of adding disjunctive clauses).

Most of the difficulties of such a system lie in the construction of suitable slave programs. Because these would be the object of manipulation by the master program it is essential that they should have a simple, regular structure, which would lend itself to easy correction. On the other hand, since the slave program is responsible for the actual translation, its structure cannot be too primitive.

Our present work on the formal approach is concerned with the development of a mathematical theory of translation which will form the basis for the construction of programs of this type. Essentially, the theory involves the description of a language model and an automaton model through which the translation process can be defined. An account of our progress so far will be found in the bibliography.

REFERENCES

1. Report on the State of Machine Translation in the United States and Great Britain, Technical Report No. 1, U. S. Office of Naval Research Information Systems Branch, Jerusalem, Israel, February 15, 1959.

2. Kaplan, A., A Study of Ambiguity and Context, Mimeographed, 18 pages, November 30, 1950; reprinted in Mechanical Translation 2:2, 39-46 (November 1955).

BIBLIOGRAPHY

Fillmore, Charles, and Koutsoudas, A., The Directed Graph in Language Description, ozalided, 29 pages, Jan., 1959.

Koutsoudas, Andreas, Mechanical Translation and Zipf's Law, Language," 33:4 (Part 1), December 1957, pp. 545-552.

Koutsoudas, Andreas, The Plural Number of Nouns, "Lang. and Speech," Vol. 1, Part 4, October-December 1958, pp. 265-268.

Koutsoudas, Andreas, Research in Machine Translation 1: General Program, Project Michigan Report, 2144-268T, March 1959.

Koutsoudas, Andreas, Defining Linear Context to Resolve Lexical Ambiguity. "Lang. and Speech," Vol. 2, Part 4, October-December 1959, pp. 211-235.

Koutsoudas, Andreas, and Halpin, A., Research in Machine Translation II: Russian Physics Vocabulary with Frequency Count, Vol.1 Right to Left Alphabetization, Vol. 2, Left to Right Alphabetization Project Michigan Report, 2144-312T, August 1958.

Koutsoudas, Andreas, and Humecky, A., Ambiguity of Syntactic Function Resolved by Linear Context, "Word," 13:3, December 1957, pp. 403-414.

Koutsoudas, Andreas, and Humecky, A., Linguistics and Machine Translation, "Word," Vol. 15, No. 3, Dec., 1959, pp. 489-491.

Koutsoudas, Andreas, and Korfhage, R., Mechanical Translation and the Problem of Multiple Meaning, "Mech. Trans.," 3:2, November 1956, pp. 46-51.

Koutsoudas, Andreas, and Korfhage, R., The Computer as a Translator, "Mich. Alumnus Quart. Review," 63:10, December 1956, pp. 34-37; reprinted in "IRE Student Quart.," 3:4, May 1957.

Koutsoudas, Andreas, and Machol, R., Machine Translation Work and the University of Michigan, "Mech. Trans.," 3:2, November 1956, p. 34.

Koutsoudas, Andreas, and Machol, R., Frequency of Occurrence of Words: A Study of Zipf's Law with Application to Mechanical Translation, Project Michigan Report, 2144-147T, June 1957.

Lehiste, Use, Order of Subject and Predicate in Scientific Russian "Mech. Trans.," 4:3, December 1957, pp. 66-67. 1

Smoke, William and Dubinsky, E., A Program for the Machine Translation of Natural Languages, ozalided, 34 pages, June, 1959.