[*Mechanical Translation*, vol.3, no.1, July 1956; pp. 20-25]

Preprogramming for Mechanical Translation

R. H. Richens

TRANSLATION is a species of communication in which the set of symbols adopted by the communicator is changed into another set of symbols before reception. It is possible to argue that all communication involves such a substitution of symbols and that communication within a single language is merely a limiting case of translation. For present purposes, however, we shall confine the scope of discussion to translation between different spoken or written languages.

We have next to inquire as to what remains invariant in translation. If we try to convey the maximum significance of the symbols of the base language, it is clear that a great deal is involved: gross meaning, the subtler overtones, deliberately concealed meanings, manifestations of the subconscious mind, the sound of the base words or their appearance in script, metrical characteristics, etymology, the associations engendered by the communication, the statistical characteristics of the communication as a sample of the output of a particular author or period, and the pleasure or otherwise engendered by communication in an informed or cultivated recipient. It is obvious that a mere fraction of all this comes over in any translation and hence we derive the notion of translation as a scaled process. We translate at various levels and in respect of various characteristics. An additional limitation on the precision of translation is provided by the peculiarities of the target language which may contain no symbol for an idea in the base language, a frequent occurrence in the case of exotic plants or animals, or no method of rendering an idea without adding an inaccurate qualifier, as in Chinese-to-English translation where the neutrality of the Chinese noun with respect to number cannot be preserved.

The notion of level or mode of translation is important. Machine translation has earned a certain notoriety for its indulgence in very low-level translation and its fondness for what has come to be known as mechanical pidgin. For certain purposes, however, such as locating allusions, low-level translation may be all that is required. Confusion only occurs if the mode of translation is not made clear.

We are now in a position to discuss the notion of preprogram. Machine translation depends on collaboration between linguists, engineers and an obscure set of people interested in the bridge territory between the two, where problems of logic and semantics arise. It is not to be expected that a person whose primary interests are linguistic will appreciate the nicer details of electronic circuitry. It is therefore important to develop procedures that are comprehensible to linguists and engineers alike and can be used as the basis for developing detailed programs for any particular machine. Such general procedures are referred to here as pre programs. Till now, the devices principally used for experiments in machine translation have been punched-card machines and electronic computers. It is possible that the best machine for machine translation as regards both efficiency and expense has not yet been devised. It is important therefore to develop procedures that are not tied down to any particular machine but which can easily be applied to a particular machine when required.

A question that is of considerable interest is the optimum combination of man and machine. It has come to be generally recognized that machine translation with intensive human pre-and post-editing is hardly worthwhile since this method is largely concerned with remedying the defects of the machine. A far more satisfactory concept is that of companionship. An efficient translating machine that can operate whenever required, can continue when its human partner is fatigued, can instruct its partner without the wearisome labor of consulting dictionaries and grammars, and can retire quietly into the background when the human partner desires to exercise his powers unaided qualifies in considerable measure as a good companion.

After these preliminaries, we can proceed directly to concrete problems.

The following convention will be used. A term in single quotes is used to represent the word in the target language of which the quotation is a common meaning.

For purposes of machine translation it is convenient to distinguish between the following operations:

1. Transfer of meaning.
2. Transfer of ambiguity.
3. Transfer of structure.
4. Injection when, for example, number is attached to a neutral Chinese noun.
5. Restraint, preventing the machine from excessive semantic analysis.

The first stage in machine translation is character recognition. There are three possible methods:

1. Complete human recognition in which a reader deals with a familiar script.
2. Incomplete human recognition in which certain visual characteristics of an unknown script are picked out.
3. Photoelectric recognition, using standard fonts.

This stage is of very considerable importance as far as the economics of machine translation is concerned, but is irrelevant to the subsequent operations and is therefore excluded from the preprogram.

The outcome of recognition is the conversion of the symbols of the base text into a functional equivalent such as holes in punched cards or teleprinter tape. Having obtained a functionalized text, the next stage is matching against a mechanical word-dictionary. This operation has been discussed in some detail by R.H. Richens and A.D. Booth[1], and I shall only refer to essentials now. Each word of the base text must be matched against the entire mechanical dictionary, searching backwards. In some cases, a presorting of the base text into alphabetical order will expedite this operation. Then, as soon as a dictionary word is encountered which is wholly contained in the base word, the equivalent or equivalents in the target language must be entered. Should there be a residue, i.e., if a base word is inflected, the residue must then be matched against the mechanical word-dictionary in its turn. In the Chinese sentence studied by the Group, affixes do not come into the picture.

A point not sufficiently considered in the earlier paper concerns languages such as Latin with different conjugations and declensions or like Welsh with initial mutation. In this case, when transferring an affix, or in Welsh, the body of the word after cutting off the mutable initials, an indication of the conjugation must be extracted from the mechanical word-dictionary. Then, when matching the detached component, the conjugation indicator must be matched simultaneously.
Thus Welsh nhroed will be decomposed into

nh	(t declension)	— no meaning
roed	(t declension)	— 'foot'

The result of this operation is the sequence of equivalents dubbed mechanical pidgin.

Matching against the mechanical word-dictionary, however, cannot be confined to the matching of single words. In most languages, irreducible compounds occur such as "cool off" which in contrast to "im-possible" cannot be analyzed into semantic components. Such irreducible compounds must be entered as such in the mechanical dictionary. Then, when matching a word which may be part of an irreducible compound, it is necessary to extract both the meanings in isolation and the meaning in combination . A second matching is then necessary to ascertain whether the other component of the potential compound is present. If this is not, the compound can be erased. If the other member of the compound is present, it may be possible to accept the compound without further operation. In the Chinese sentence under consideration, the chances of encountering yung[2]-chieh[3] 'dissolve' in which the components retain their isolated meanings are relatively low.

It may be necessary, however, as in the case of German separable verbal prefixes, to defer a decision as to whether an irreducible compound is present until the syntax has been analyzed.

Whenever a compound is accepted, the meanings of the components in solution must be erased.

Thus, to obtain an output in mechanical pidgin, the mechanical dictionary must contain the words or parts of words of the base language, irreducible compounds, the equivalents in the target language, and indications of conjugation. In order to translate at a higher level, a more elaborate mechanical dictionary is required.

There are two types of information that we can utilize at our next level, syntactical and semantic. In the sentence "the dog bites the cat", subject and predicate are distinguished syntactically; in the sentence "this plant has yellow petals", semantic analysis indicates a botanical rather

1. Machine Translation of Languages. New York 1955, p. 24.

Sentence	WC	R	A	Pre	Post	WC	R	A	Pre	Post	WC	R	A	Pre	Post	WC
'however'	Advs	1				Advs	1				Advs	1				
'this'	Dem	2	2			NPp	2				NPp	2				
'two'	num	3				of										
'entity'	N^2	4	2													
'of'	of	1														
'appearance'	N^1	1		the		NPp	1									
'and'	and	2	1													
'dissolve'	V^2	4		of	-ing											
'degree'	N^1	3														
'somewhat'	Adv	1				Adj	1				Adj	3		are		
'not'	neg	2														
'alike'	Adjv	3	1													

WC -- word class
R -- rearrangement
A -- alternative
Pre -- preinsertion
Post -- postinsertion
Adjv -- adjective adverb
Advs -- adversative
NPp -- plural noun phrase

Table 1 : Schedule for syntax analysis

than engineering significance for "plant". Syntactic information will be dealt with first since it appears to present rather less complex problems than semantic information.

In order to analyze syntax, it is convenient to allocate words to word classes. In some cases these can be parts of speech or parts of speech delimited in various ways. Sometimes, in the Chinese chi[2] 'and', in which "reach" is an alternative meaning, the word class will be the sum of "and" and "verb". There is nothing against using different categories of word classes for different pairs of languages, though a general unified scheme has some obvious advantages. It is useful to allocate some of the most frequent multipurpose words to one-member classes of their own.

For utilizing syntactical information the mechanical dictionary must contain expressions for the word class of each entry; this will take the form of a number or series of numbers for each word. When translating at this level, the preliminary matching process now results in the output of a sequence of word class expressions corresponding to the sequence of words in the base text. There are now various possibilities. Dr. Parker-Rhodes would use the word classes to provide material for a computing schedule based on a moderately restricted set of instructions. I take this as analogous to learning a foreign language by means of a grammar. The method suggested here is more analogous to learning one's native tongue, in which correct usage is arrived at by imitation over a long period with no conscious realization of rules.

The mechanical dictionary in the present method must contain a supplementary dictionary of word-class sequences. The sequence of word classes for a single sentence is then treated as a single compound or inflected word. This is decomposed into its constituents in the same way as the individual words are decomposed into stem and affix, that is by matching the initial component first and then proceeding to the next and so on to the end. It is possible that, in the case of word-class sequences, the front may not be the best place to start, at least in some cases. This is a matter for further investigation.

The mechanical word-class sequence dictionary contains the following data under each entry:

1. Word-class sequence.
2. Rearrangement instructions.
3. Alternative instructions.
4. Pre- and post- insertion instructions.
5. Word-class equivalent.

The result of the matching procedure against the word-class sequence dictionary is to generate a series of instructions and a new word-class sequence. The latter then provides the basis fora new cycle of matching against the word-class sequence dictionary. The whole procedure is repeated until a word-class sequence is generated that is wholly contained in the mechanical dictionary. The operation is then concluded.

The accumulated instructions can then be read off, the rearrangements made, alternatives eliminated, and the necessary insertions made. In the Chinese sentence, three reductional cycles were involved. The procedure is illustrated in Table I. The output reads "however the appearance and degree of dissolv- ing of these two entities are somewhat un- alike".

The information utilized so far has been syntactical. The semantic information is more difficult to process and what follows is merely tentative.

A possible method is to attach semantic indicators to significant words and to collect the indicators as one proceeds through a passage, using the totals to decide between alternative renderings of doubtful words. Thus "petal", "stem" and "pineapple" could be accompanied by indicators for "botanical". This might help to limit "plant" to its botanical rather than its engineering sense. As Dr. Thouless has pointed out, some difficulty might be encountered with a "pineapple-slicing plant", but in this case "slicing" might carry an indicator pointing the other way. I am not in a position to say how useful this method could be. It has the advantage of collecting information as the text is traversed. However, it is obviously an extremely crude way of mobilizing semantic information and I should therefore like to consider next a more difficult but more fundamental approach.

I refer now to the construction of an interlingua in which all the structural peculiarities of the base language are removed and we are left with what I shall call a "semantic net" of "naked ideas". These bear some obvious resemblances to the linguistic configurations discussed already.

The elements represent things, qualities or relations. I associate adjectives (usually monadic relations) and verbs (dyadic or higher relations) in the Japanese way.

A bond points from a thing to its qualities or relations, or from a quality or relation to a further qualification.

"black cat" is

cat ⟶ black

"The cat is on the mat" or

"The mat is under the cat" is

cat $\xrightarrow{1}$ on $\xrightarrow{2}$ mat

In asymmetrical relations, the bonds are not interchangeable.

"The dog bites the cat" can be represented as

dog $\xrightarrow{1}$ part of $\xleftarrow{2}$ teeth $\xrightarrow{1}$ contact $\xleftarrow{2}$ cat

contact ↓ much

If a different category of bond is used for doubtful or uncertain connections, a method of precisely delimiting the field of ambiguity is available.

Constructions of the type dog ⟶ part of ⟶ teeth are not used since this would assume the possibility and desirability of weighting the terms of dyadic relations in terms of "superiority" or "inferiority".

When the Chinese sentence studied by the Group is represented as a semantic net, the figure obtained is of considerable complexity. What is more, various deficiencies in the information provided by the sentence become apparent; for instance, no mention is made of the solvent, without knowledge of which the significance of "solubility" is vacuous.

This raises the question of "restraint". A translator is frequently under the necessity of reproducing ambiguities or inconsistencies in the base language by corresponding ambiguities or inconsistencies in the target language. If a machine is to utilize semantic data, it must necessarily analyze the semantic relations of the passage fed into it. If this analysis is carried too far, the base passage is in danger of such severe mangling that a readable output in the target language will not be obtained. Thus in the example quoted, a machine that indulges in semantic analysis will demand information on the solvent; if however, it is restrained to conform to the frailties of human nature, it should be possible to stop analysis at the level of the concept "solubility" and present the smooth inadequate output that a human translator is expected to provide. It might prove possible to arrange for a machine to translate at various levels of restraint so that the ordinary person and the logician can each be satisfied.

The semantic net thus represents what is invariant during translation. It can, of course, be transformed into a unique linear sequence for dictionary purposes, rather in the way that the structural formulae of organic compounds can be given linear codes for purposes of cataloguing.

The problem of extracting semantic nets from base texts is difficult and no general mechanical procedure has yet been devised. One possibility is to regard the words of the base passage as pieces in a jigsaw puzzle. Each word has a number of semantic properties - differently shaped protuberances in the jigsaw analogy - which fit in with some words but not with others.

Thus the relation " $\xrightarrow{1}$ see $\xleftarrow{2}$ " can only attach on the left-hand side to a human being or animal. Syntax already restricts the number of possible combinations; semantics limits the possibilities still further.

If syntax and semantics do not lead to a unique interlocking, we have an ambiguous situation. Ambiguity can be represented in a semantic net by introducing a second category of bonds, and can presumably be transferred to the target passage if so required.

The syntactical procedure discussed earlier in this paper dealt with a specific pair of languages. It is more satisfactory theoretically to go through an interlingua that is capable of expressing the nuances of all the languages considered in a translation program and is more adequate for logical analysis than any existing language. Such an interlingua would have the practical advantage of connecting such languages as Welsh and Japanese, where the labor of compiling a specific translation program would not be worthwhile. It is well known that two-stage translation via an intermediary language is unsatisfactory; this is only so, however, when the intermediary language is a natural rather than a universal language.

The semantic nets described above have an obvious bearing on the question of a universal interlingua. If the elements (ideas) are replaced by letters with an ideographic significance only, we have in fact an ideographic algebraic script with obvious potentialities for machine translation work. The elaboration of a system of ideographs for handling discourse is one of the current research projects of the Cambridge Group.

In conclusion, I would like to return to the notion of translation as a scaled process in which a selection has to be made of the amount of information to be transferred. It is only a further

step to the notion of translation as a limiting case of abstracting. In ordinary academic life, especially in science, abstracts are required far more frequently than full translations. In the future, the increased rate of publication is likely to make the production of abstracts far more necessary. It therefore seems that any procedure of selective transfer of ideas is likely to be of considerable future interest. Semantic nets have an obvious relevance in this connection. This paper had, as its object, a brief description of some of the work being done by the Cambridge Language Research Group on machine translation. This work has now reached the stage where one is beginning to dabble seriously in schemes for machine abstracting.