[*Mechanical Translation*, vol.4, nos.1 and 2, November 1957; pp. 11-13]

Semantic Frequency Counts

Paul Pimsleur, University of California, Los Angeles, California

The success of a mechanical translation should be measured in terms of the level of depth required by the situation. To determine whether a careful translation is desirable a rough scanning will suffice. The use of cover-words, high frequency words that may be substituted for low frequency words, in the output language is an essential part of this process. The preparation of trans-semantic frequency counts resulting in dictionaries of reduced size that require less computer storage capacity is recommended.

ACCORDING to Y. Bar-Hillel, "The central problem in mechanizing translation is the preparation of methods that permit a more restricted memory. Hitherto accepted methods require a rapid access mechanical memory with storage capacity greatly in excess of that of available electronic computers."[1]

Though work is now in progress on machines featuring large density storage units and rapid access time, [2] the development of such machines will not substantially change the problem. The goal is, and will remain, the creation of the most efficient dictionary for MT purposes, containing the smallest number of entries and featuring the most rapid search procedures.

The reduction of dictionary size is directly related to the matter of multiple -meaning. The ideal dictionary will be the smallest possible one which still suffices to meet the requirements of translation, within the limits of accuracy we have chosen to accept. However, such a dictionary presupposes considerable knowledge of the frequency with which words occur, in each of their several meanings. "In effect, what is needed are true ideoglossaries, based on actual, rather than potential, behavior."[3] Though some attempts have been made to attack this problem as it has arisen in particular research contexts, [4] no concentrated effort is being exerted toward the establishment of semantic frequency counts per se. It appears, however, that such counts are essential to the future development of MT. Some additional incentive may also be derived from the recent indications that Russian MT specialists have been working for some time on a "polysemantic dictionary" which is a central part of their MT procedure.[5]

A semantic frequency count is a listing of the words of a language, with the several meanings of each word, and the relative frequency of occurrence of each meaning in general and/or specialized contexts. Valuable as such a count might be to scholars and educators in various domains, it appears that a somewhat different count is needed for purposes of MT. The need is for TRANS-SEMANTIC FREQUENCY COUNTS. A trans-semantic frequency count is a listing of the words of the source language, together with the various possible renderings of each in the target language, and the frequency of occurrence of each of the latter. Such a listing would resemble a normal translation dictionary, with the addition of information, probably in the form of percentages, giving the

---

1. Y. Bar-Hillel, "Can Translation be Mechanized, " (abstract) MT, Vol.3, No. 2, p. 67.

2. G.W. King, "Stochastic Methods of Mechanical Translation, " MT, Vol. 3, No. 2, pp. 38-39.

3. K.E. Harper, "Contextual Analysis in Word-for-Word MT, " MT, Vol.3, No. 2, p. 40.

4. A. Koutsoudas and R. Korfhage, "Mechanical Translation and the Problem of Multiple Meaning," MT. Vol.3, No. 2, pp. 46-51, 61.

5. D. Panov, "On the Problem of Mechanical Translation, " MT, Vol.3, No. 2, pp. 42-43.

frequency of occurrence of each meaning in the target language. Alternate frequencies should also be given for various subject areas, scientific, military, etc.

As described here, such an undertaking would be enormous, even for any two languages. However, it may be argued that: 1) the need for such information is great for MT; 2) any partial listing would provide data that could immediately be useful in the preparation of MT dictionaries.

In connection with the problem of multiple-meaning, it may be useful to dwell briefly on another approach. Virtually all non-mechanical translators, and even some who are concerned with MT, think in terms of sure translation. By sure translation is meant a sort of one-to-one semantic mapping from the words of the source language to the best possible "mots justes " of the target language. The suggestion is offered that the issue be rephrased in terms of probabilities (a "stochastic approach"[6]), in which we aim at the degree of success in translation which the situation seems to demand. By success is meant a comprehensible, non-misleading rendering. The degree of success may well vary with the danger or inconvenience resulting from imperfect translation. In many instances, there may be quantities of material to be merely scanned for purposes of determining whether any use is to be made of any part of it. In such cases, a very rough translation has been shown to suffice,[7] with a consequent saving in cost and intricacy of machine operation. A minimum probability coefficient of .80 for each ambiguous word may be sufficient for such rough scanning. This sort of translation is probably attainable in the relatively near future, though anything like a "perfect" translation is still on the distant horizon.

Thus the concept of levels of depth becomes important. The first level of depth may be a translation in which the chances are 80 or more out of a hundred that each ambiguous word has been translated acceptably. The second level of depth might involve a minimum confidence of 90% per word; the third and most refined level (the one on the distant horizon) would provide confidence .95 or perhaps even .99 per multiple-meaning word. This concept may be symbolized as:

$$\Pr(X \text{ is acceptable}) \geq 1-\alpha$$

where Pr means "the probability that. .. ", X represents a given rendering of a source word in the target language, and a stands for the maximum tolerable error per word. In the levels of depth just discussed, the alphas would be .20, .10, and .05 or .01, respectively. Obviously, each successive level will require considerably more search-time, an improved and probably a larger dictionary, and more detailed programming.

An illustration may serve to clarify several concepts. In the German sentence

Die Aufgabe ist zu schwer.[8]

the word schwer presents a typical problem in multiple-meaning. A dictionary of modest dimensions [9] lists the following eight meanings, for each of which we have provided an English translation. (Several sub-meanings listed as colloquial have, perhaps unfairly, been omitted.)

1) 'weigh-s' (verb). Die Kiste ist drei Zentner schwer, 'the box weighs three hundredweight .'
2) 'heavy'; 'strong.' ein schwerer Stein, 'a heavy stone;' ein schwerer Wein, 'a strong (intoxicating) wine.'
3) 'laden.' Das Dach ist schwer von Schnee. 'the roof is laden with snow.'
4) 'difficult.' Das fällt mir schwer, 'I find 'that difficult.'
5) 'unfortunate'; 'hard.' Er hat ein schweres Schicksal, 'he has an unfortunate fate.' Sie nimmt es schwer, 'she takes it (the news) hard.'
6) 'very.' Der Mann ist schwer reich, 'the man is very rich.'
7) 'slow-ly.' Er ist schwer von Begriff, 'he is slow to catch on,' or 'he catches on slowly.'
8) 'pregnant.' Die Lage ist schwer an Entscheidungen, 'the situation is pregnant with decisions.'

---

6. G. W. King, "Stochastic Methods of Mechanical Translation," MT. Vol.3, No. 2, pp. 38-39.

7. J.W. Perry, "Translation of Russian Technical Literature by Machine, " MT. Vol. 2, No. 1, (discussion of results) p. 16.

8. T.M. Stout, "Computing Machines for Language Translation, " MT, Vol. 1, No. 3, p. 41.

9. Der Sprach-Brockhaus. Eberhard Brockhaus, Wiesbaden, 1954.

There are thus ten possible translations for the German word schwer, in this no doubt incomplete list. They are: 'heavy, strong, laden, difficult, unfortunate, hard, pregnant, slow-ly, very, weigh-s.' By introducing the concept of COVER-WORDS, the number of these translations can be substantially reduced.

A cover-word is a word of relatively high semantic frequency which can be used in place of words of lower semantic frequency, with little possibility of misinforming the reader.

Referring back to the list above, let us examine each of the meanings of schwer in turn.

1) 'weigh-s' (v.i.) requires the translation of a predicate adjective in German by a verb in English — though these grammatical concepts may be operationally meaningless in MT, they are retained here for convenience. The importance of the problem depends on the frequency of occurrence of this locution, which is unknown at present. A trans-semantic frequency count would help us to decide how situations of this sort are to be handled. In any event, the possibility should be considered of using the awkward translation, 'the box is three hundredweight heavy,' thereby using the cover-word 'heavy' for 'weighs.' The loss is primarily of elegance, not of correct understanding.

2) 'heavy' needs no comment; it is a primary, or high-frequency rendering. 'Strong' would seem to be infrequent enough to render it inconsequential, but this again must be confirmed empirically.

3) 'laden.' If we rendered 'the roof is laden with snow' by 'the roof is heavy with snow,' the cover-word is used and no misinterpretation can result.

4) 'difficult' is a high-frequency meaning and appears irreduceable. This again must be checked empirically, which presupposes a trans-semantic frequency count.

5) 'unfortunate' may be replaced by 'heavy' in the sentence 'he has a heavy fate,' with a loss of elegance but little semantic distortion. The meaning 'hard,' as in 'she takes it hard' is somewhat more troublesome. Whether it is worthwhile to program special instructions for dealing with this case will depend on the frequency with which it can be expected to occur. In scientific literature at least, the frequency may be negligible. Should special provision for this case be necessary, it might be best to treat it as a compound, etwas schwernehmen.

6) 'very.' Schwer reich should be translated as 'very rich,' while schwer verletzt means 'badly wounded,' and schwer enttäuscht may be either 'badly disappointed' or 'very disappointed.' The solution seems to lie in translating schwer in this context as 'very,' thus forcing acceptance of 'he was very wounded' instead of 'he was badly wounded.' It appears necessary to allow 'very' as a third rendering of schwer, alongside 'heavy' and 'difficult.' However, its occurrence as 'very' may be limited to cases such as those cited above, where it is directly followed by one of a small number of adjectives and can thus be identified rather easily by the machine.

7) 'slow-ly.' Schwer von Begriff requires special treatment as an idiom.

8) 'pregnant' can be rendered by the cover-word 'heavy' without serious loss.

Thus the ten meanings of schwer have been reduced to three cover meanings, 'heavy, difficult and very,' of which only 'difficult' and 'heavy' may be expected to occur in many different settings which we cannot at present predict. No loss of comprehension has resulted from the use of cover-words, though stylistic violence has been done to a varying extent. This drawback is offset by a substantial gain in terms of machine time and storage space.

SUMMARY AND CONCLUSIONS

1. It has been suggested that work be undertaken with all possible speed toward the establishment of trans-semantic word counts, with the goal of attaching a probability coefficient to the occurrence of a given meaning of a given word in a given subject field. Without underestimating the enormousness of the task, it is submitted that it is indispensable to MT. The work should commence with the subject areas of most immediate concern, i.e. scientific, and with the words which occur with greatest frequency, as shown by existing word-counts of the major languages. New machine methods may lighten the task considerably.

2. The concept of levels of depth has been used to describe translations of differing (but predictable) degrees of accuracy.

3. The concept of cover-words has been used, as well as that of trans-semantic frequency counts, to assist in reducing the contents of a storage dictionary.