[Mechanical Translation, vol. 8, No. 2, February 1965]

A Figure of Merit Technique for the Resolution of Non-Grammatical Ambiguity

by Swaminathan Madhu, General Dynamics/Electronics, Rochester, New York, and Dean W. Lytle*, University of Washington, Seattle, Washington

Ambiguity in language translation is due to the presence of words in the source language with multiple non-synonymous target equivalents. A contextual analysis is required whenever a grammatical analysis fails to resolve such ambiguity. In the case of scientific and engineering literature, clues to the context can be obtained from a knowledge of the varying degrees of probability with which words occur in different fields of science. A figure of merit is defined, which is calculated from the probability of word occurrences, and which leads to the choice of a particular target equivalent of a word as the most probably correct one. The results of applying the technique to a set of twenty one Russian sentences indicate that the technique can be successful in about 90% of the cases. The technique can easily be adapted for use by a computer.

Introduction

Ambiguity in automatic language translation is due to the presence of words in the source language with more than one equivalent in the target language. The elimination of such polysemantic ambiguity is essential in order to make the translation readable and useful. Polysemantic ambiguity may broadly be classified into two types: one in which grammatical processing can be used effectively to get rid of the superfluous target equivalents, and the other in which grammatical processing is ineffective. We confine ourselves here to the latter type of ambiguity, the non-grammatical ambiguity.

The resolution of non-grammatical ambiguity requires some kind of contextual analysis; and, in the case of mechanical translation, the contextual analysis should be such that it can be readily performed by a computer.

A method for the automatic resolution of non-grammatical ambiguity was reported in 1958 by the MT group at the University of Washington.[1] According to that method, a field of science classification scheme was used in which the entire area of science and engineering was divided into nearly seventy fields of science. A few of the words in the target language were then tagged with numbers representing the particular field of science in which they occurred almost exclusively. Since the number of words that could be tagged in the above manner was small, the method was found to be successful only in a very small number of cases to which it was applied.

This paper uses the field of science classification scheme mentioned above as a starting point, but approaches the problem of non-grammatical ambiguity from the viewpoint of probability theory. A "figure of merit" technique is developed which promises to be highly effective in the translation of scientific and engineering literature.

The Basis of the Figure of Merit Technique

When the occurrence of a multiple meaning word, i.e., a source language word with more than one target equivalent, causes non-grammatical ambiguity, the appropriate target equivalent can be chosen by an examination of the context in which the multiple meaning word occurs. For example, the Russian word *uzlov* has the following English equivalents*: 'knots', 'junctions', 'bundles', 'nodes', 'assemblies', 'ganglia', and 'joints'. If the word *uzlov* occurs in an article discussing the central nervous system of the human body, the correct choice is probably 'ganglia'. On the other hand, if it occurs in an article on electrical network analysis, the appropriate choice is 'nodes'. In these examples, the context is determined by noting the particular branch of science to which the article belongs. Such a criterion is evidently most useful in the case of scientific and engineering literature. When the article cannot be clearly classified as belonging to a specific scientific field, the determination of the context must be made on a probabilistic basis.

The figure of merit technique is based on the premise that context can be determined by a consideration of

* The authors wish to thank Dr. David L. Johnson, Department of Electrical Engineering, University of Washington, for many valuable suggestions and discussion of the material in this paper. This work was supported by a contract from the U.S. Air Force, Rome Air Development Center, and this help is gratefully acknowledged.

[1] University of Washington, *Linguistic and Engineering Studies in Automatic Translation of Scientific Russian into English,* Department of Far Eastern and Slavic Languages and Department of Electrical Engineering, University of Washington, Seattle, Washington, 1958.

* The English equivalents of the Russian words cited in this paper will be those listed in the dictionary compiled by the MT group at the University of Washington, Seattle, Washington.

the probability of occurrence of a given target equivalent in a particular field of science. The frequency with which a target equivalent occurs in one field of science is, in general, different from that in another field of science. A few target equivalents occur almost exclusively in one field of science; e.g., the phrase 'blue-green algae' is encountered most often in the area of biological sciences. The vast majority of target equivalents, however, occur in several different fields of science, but with a different probability of occurrence in each of them. The figure of merit tries to take advantage of the different probabilities of occurrence of a word in different fields of science. It is possible to determine the probability measures of a sufficiently large number of target equivalents by means of a statistical analysis, as will be described in the next section.

The underlying principles of this method will now be considered. In any article being translated, there are multiple meaning words as well as words with single target equivalents. The latter will be called "single meaning words" for the sake of simplicity. The target equivalents of the single meaning words have different degrees of probability of occurrence in the different fields of science. Therefore, an examination of the single meaning words found in an article along with their probability measures, will provide a clue to the context in which the multiple meaning words occur in the same article. For instance, if the article being translated deals with a mathematical topic, then the single meaning words occurring in it will generally have a higher probability of occurrence in mathematics than in other fields of science. Therefore, by operating upon the probability measures of single meaning words found in an article, the context in which they occur can be estimated.

When the context has been determined in this manner, the most probably correct target equivalent of each multiple meaning word can be chosen so as to conform to the context. This again will require suitable operations on the probability measures of the several target equivalents of a multiple meaning word, so that these measures will be correlated with the context.

Collection and Organization of Data on Word Occurrences

In order to assign relative probability measures to a fairly large number of target equivalents, a statistical analysis was performed manually on a collection of 111 Russian texts* (and their English translations) dealing with a multitude of scientific topics. In the analysis, use was made of the word-for-word translations retaining all the allowed target equivalents of Russian multiple meaning words, as well as the "free" translations in which the ambiguity had been resolved by a human translator. Since the aim was to eliminate non-grammatical ambiguity, words such as prepositions, the definite and indefinite articles, were ignored. Moreover, very common words as, for example, the verb 'to be' and its various forms, that occur indiscriminately in the literature of all branches of science were also ignored, since they provide no clue to the context. Only the remaining words and their occurrences were noted in the analysis.

The entire area of science and engineering was subdivided into nearly seventy sub-fields of science, e.g., optics, acoustics, biochemistry, etc.* Each paragraph of the Russian texts was classified according to the sub-field of science to which it belonged. For each of the English words occurring in the translations (with the exceptions mentioned earlier), a count was made on how often it occurred in the different sub-fields of science. In this analysis, data on the relative frequencies of occurrence were collected for 3400 different English words with a total number of occurrences equal to 14385.

In order to organize the data collected, the entire set of nearly 70 sub-fields of science was rearranged into ten large groups. This regrouping was necessary since the original classification contained far too many different fields, and the use of nearly 70 sub-fields made too fine a distinction between related sub-fields of science. The formation of ten large groups took into consideration the inherent similarity in the basic vocabulary of several different branches of science. Several fields of science could be grouped together on the basis of their having a large number of words common among themselves. The number of groups was arbitrarily fixed at ten. The contents of the ten groups were as follows:

Group I: Mathematics, Physics, Electrical Engineering, Acoustics, Nuclear Engineering;
Group II: Chemistry, Chemical Engineering, Photography;
Group III: Biology, Medicine;
Group IV: Astronomy, Meteorology;
Group V: Geology, Geophysics, Geography, Oceanography;
Group VI: Mechanics, Structures;
Group VII: Mechanical Engineering, Aeronautical Engineering, Production and Manufacturing Methods;
Group VIII: Materials, Mining, Metals, Ceramics, Textiles;
Group IX: Political Science, Military Science;
Group X: Social Sciences, Economics, Linguistics, etc.

On the basis of the above groupings and the data on word occurrences, it was possible to calculate the probability measures of 3400 English words.

* Each text was a part of an article dealing with some scientific subject and consisted, on the average, of about twenty sentences.

* This subdivision was originally carried out by Professor W. Ryland Hill of the Department of Electrical Engineering, University of Washington.

Probability Measures of Target Equivalents

The three probability measures that are of importance here are: (a) conditional probability; (b) marginal probability; (c) joint probability.

The conditional probability used here represents the probability of having a certain group (I, II, . . ., X), given that a particular target equivalent W_k occurs. This is denoted by the symbol $p(N/W_k)$, where N represents the group number, N = I, II, . . . , X. The conditional probability is calculated from the equation:

$$(1)\ p(I/W_k) = \frac{\text{(Number of times the target equivalent } W_k \text{ occurred in Group I)}}{\text{(Total number of times the target equivalent } W_k \text{ occurred in the entire analysis)}}$$

Similar relations are used for calculating $p(II/W_k)$, $p(III/W_k)$,etc.

The marginal probability measure used here represents the probability of having the target equivalent W_k regardless of what group it occurred in, in the entire analysis. This is denoted by the symbol $p(W_k)$, and is given by

$$(2)\ \ p(W_k) = \frac{\text{(Total number of times the target equivalent } W_k \text{ occurred in the entire analysis)}}{\text{(Total number of word occurrences in the entire analysis)}}$$

Since the total number of word occurrences in the analysis was 14385, the denominator of equation (2) could be replaced by this number. These values of $p(W_k)$, however, tended to be inconveniently small, and resulted in rather involved bookkeeping of the correct number of decimal places in the various calculations. Consequently, a scale factor was introduced so as to make the smallest value of $p(W_k)$ equal to 0.1, i.e., each value of $p(W_k)$ was multiplied by a factor of 1438.5.

In view of the scale factor introduced, the adjusted values of $p(W_k)$ are not strictly marginal probability measures in a precise mathematical sense. They will, therefore, be called "marginal *frequency* measures" in the following discussion. For the same reason, the term 'joint *frequency* measure' will be used here instead of 'joint probability measure', to represent the probability that the target equivalent W_k and the Group N have occurred together. The joint frequency measure of the combined occurrence of the target equivalent W_k and the Group N is denoted by $p(W_k,N)$ or $p(N,W_k)$. The values of this measure are calculated from the conditional probability measures and the marginal frequency measures by using the equation

$$(3)\qquad p(W_k,N) = p(N/W_k)p(W_k)$$

These three quantities,—the conditional probability measure, the marginal frequency measure, and the joint frequency measure,—were calculated for the 3400 English words occurring in the sample used. These values can be operated upon so as to provide a clue to the elimination of superfluous target equivalents of multiple meaning words.

Details of the Figure of Merit Technique

The figure of merit technique uses the probability measures of the single meaning words in an article (or sentence) to obtain a measure of the context in which the multiple meaning words in that article (or sentence) occur. The probability measures of each target equivalent of a multiple meaning word are then correlated with the context to obtain a figure of merit which allows the selection of one of the target equivalents as the most probably correct meaning in the given context.

Since the method depends upon the availability of the probability measures of target equivalents, only those target equivalents for which such information is available from the data are used in the calculations described below. The method can be used to handle each sentence separately, or a set of sentences together. In what follows, each sentence will be assumed to be treated separately.

The words from each sentence of the source language text are selected, and their target equivalents along with their joint frequency measures are noted and arranged in a tabular form. The joint frequency measures of the single meaning words are added separately for each group, i.e., the values in each column for the single meaning words are added. This yields a set of ten numbers that will be called the "marginal frequency measures of the group". If p(I) denotes the marginal frequency measure of Group I, then

$$(4)\ p(I) = p(W_1,I) + p(W_2,I) + .. + p(W_k,I)$$

where it is assumed that there are k single meaning words in the sentence, and the summation is over the single meaning words only. Similar equations can be written for p (II), p (III), etc.

The simplest procedure would seem to be: (a) to find the group for which p(N) has the highest value, and classify the sentence as belonging to that group, say, Group IX; and (b) to choose that target equivalent W_m of a multiple meaning word for which $p(W_m/IX)$ is the greatest. The values of $p(W_m/N)$ could be readily calculated by using Bayes's Theorem:

$$(5)\qquad p(W_m/N) = \frac{p(N/W_m)p(W_m)}{\sum_m p(N/W_m)p(W_m)}$$

This procedure would allow the selection of the most probably correct target equivalents in a certain number of cases. Nevertheless it was not adopted for several reasons. In some sentences, no single group might have a maximum value of p(N), in which case the above procedure would be inapplicable. More importantly, the above procedure would completely ignore the influence of all but one group on the selection of the correct target equivalents, even when other groups had values of p(N) only slightly smaller than the maximum value of p (N). A more general approach seems to be one in which each group contributes a certain weight to the target equivalent being considered, and in which the target equivalent with the maximum weight is chosen as the most probably correct one. The weight contributed by each group should depend upon the marginal frequency measure of the group itself, as well as upon the joint frequency measure of the combined occurrence of that group and the target equivalent being considered. This leads to the following definition of a figure of merit of a target equivalent W_m,

$$(6) \quad \textbf{Figure of Merit of } W_m = \sum_N p(N)p(W_m,N), \quad N = I, II, \ldots$$

The calculation of the figure of merit can also be expressed in matrix notation as follows. Define a row matrix A as consisting of the ten values p(I), p(II), . . . , p(X). Define a row matrix B as consisting of the ten joint frequency measures $p(W_m,N)$ for a given target equivalent W_m of a multiple meaning word. Then,

$$(7) \qquad \text{Figure of Merit of } W_m = AB_t$$

where B_t denotes the column matrix obtained by transposing B.

The figure of merit can be calculated for each of the allowed target equivalents of a multiple meaning word, and the target equivalent with the highest figure of merit selected as the most probably correct one for the given multiple meaning word in the given sentence.

An Illustrative Example

The application of the above procedure to an actual example will be presented in this section. The "simulated"* translation of two Russian sentences occurring in an article is as follows:

SYSTEMATIZATION/TAXONOMY/ (of) SYSTEMATIST (of) - OLD BLUE-GREEN * (of)BLUE-GREEN-ALGAE MUST/ SHOULD/OWE(s) (to)BE-BASED ON/IN/AT/TO/FOR/- BY/WITH (of) MORPHOLOGICAL * MORPHOLOGICAL-FEATURES (of) REMAINDERS/RADICALS (of) SELVES (of)-PLANTS. WITH/FROM/ABOUT (by/with/as)CONSIDERATION/CALCULATION/REGISTRATION (of)STRUCTURE/-BUILDING(s) (of)ONE/ALONE (of)DOUBLE/GEMINATE (of)ANNUAL/YEARS (of)LAYER/LAMELLA (of) (to/for) (by/with/as)LINE (of)THIN-CRUST(s) HOW/AS/BUT (of) (to/for) (by/with/as)FOSSILIZED (of)(to/for)-(by/with/as) ALGAE/WATER-PLANT * (of)(to/for)-ALGAE-COLONY;

Table 1 shows the values of the joint frequency measures of the various target equivalents occurring in the above example. The bottom row lists the values of the marginal frequency measures for the ten groups obtained by using Equation (4). For example, for Group III,

$$(8) \qquad p\,(III) = 0.5 + 2.0 + 2.8 + 0.5 = 5.8$$

The figures of merit for the different target equivalents of each multiple meaning word in the sentence are calculated by using Equation (6), and the results obained are shown in the last column of Table I. For example,

$$\text{Figure of Merit of 'STRUCTURE'} = (0.1x2.6) + (1.9x5.8) + (0.8x4.2) + (0.1x1.8) + (0.3x0.5) = \underline{14.97}$$

For each multiple meaning word, the figures of merit of the different target equivalents are compared, and the one with the highest value is selected as correct.

For example, in the case of 'STRUCTURE/BUILDING', the figure of merit for 'STRUCTURE' is 14.97, while that for BUILDING' is 2.6; and the choice is 'STRUCTURE'. In Table I, the selection for each multiple meaning word is indicated by italicizing the corresponding figure of merit.

Testing the Validity of the Technique

A set of 21 sentences selected from Russian journals dealing with chemistry and with radio engineering was used to test the figure of merit technique. These sentences were unrelated to the ones used in the collection of data on word occurrences. This selection will summarize the results obtained from the test set*.

In the 21 sentences, there were a total of 202 words :hat were of interest and had their target equivalents listed in the bilingual tagged lexicon used as a reference. Of these 202 words, 76 were multiple meaning words with a total of 172 English equivalents. The figure of merit technique enabled the choice of correct equivalents for 66 out of the 76 multiple meaning words. The correctness of the choice was judged by examining the intended meaning of the original Russian sentences.

There were 10 multiple meaning words for which the target equivalents chosen by the above procedure were partly, or sometimes wholly, inappropriate. In most of these cases, the incorrectness was attributable to the fact that the source of the data on word occurrences *was* limited in size, and also biassed rather heavily in Favor of the biological and medical sciences. Consequently, target equivalents with a higher probability of occurrence in Group III were selected in some sentences

* The "simulated" translation simulates the output from a computer with all the superfluous target equivalents retained. A slash "/" between words indicates that one of the words has to be selected. An asterisk preceding a phrase indicates an idiomatic form recognized by the computer.

* A more detailed discussion and the calculations can be found in: "Translation Study: Final Report," Department of Electrical Engiicering, University of Washington, Seattle, Washington, 1961, pp. 170-229.

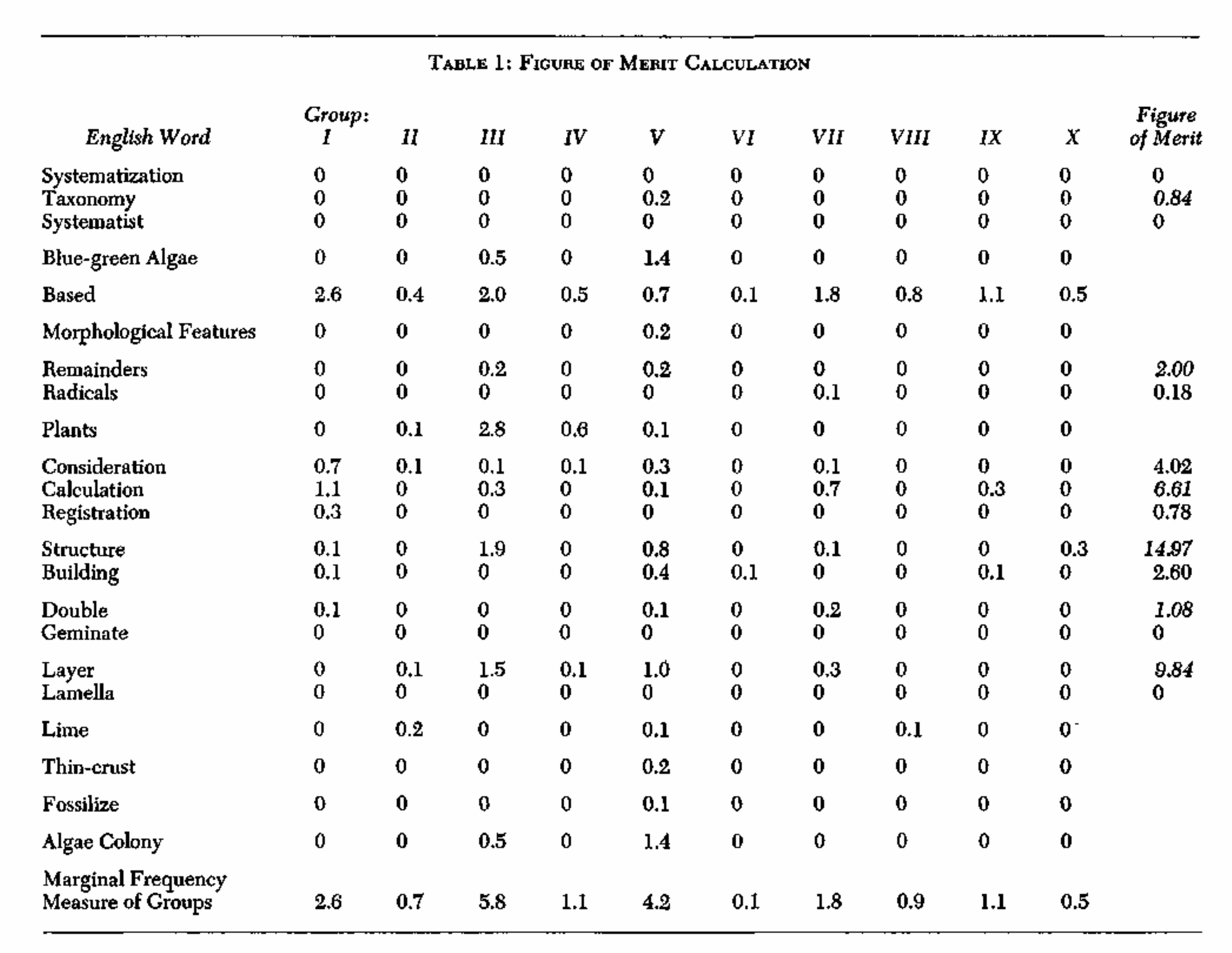

TABLE 1: FIGURE OF MERIT CALCULATION

English Word	Group: I	II	III	IV	V	VI	VII	VIII	IX	X	Figure of Merit
Systematization	0	0	0	0	0	0	0	0	0	0	0
Taxonomy	0	0	0	0	0.2	0	0	0	0	0	*0.84*
Systematist	0	0	0	0	0	0	0	0	0	0	0
Blue-green Algae	0	0	0.5	0	1.4	0	0	0	0	0	
Based	2.6	0.4	2.0	0.5	0.7	0.1	1.8	0.8	1.1	0.5	
Morphological Features	0	0	0	0	0.2	0	0	0	0	0	
Remainders	0	0	0.2	0	0.2	0	0	0	0	0	*2.00*
Radicals	0	0	0	0	0	0	0.1	0	0	0	0.18
Plants	0	0.1	2.8	0.6	0.1	0	0	0	0	0	
Consideration	0.7	0.1	0.1	0.1	0.3	0	0.1	0	0	0	4.02
Calculation	1.1	0	0.3	0	0.1	0	0.7	0	0.3	0	*6.61*
Registration	0.3	0	0	0	0	0	0	0	0	0	0.78
Structure	0.1	0	1.9	0	0.8	0	0.1	0	0	0.3	*14.97*
Building	0.1	0	0	0	0.4	0.1	0	0	0.1	0	2.60
Double	0.1	0	0	0	0.1	0	0.2	0	0	0	*1.08*
Geminate	0	0	0	0	0	0	0	0	0	0	0
Layer	0	0.1	1.5	0.1	1.0	0	0.3	0	0	0	*9.84*
Lamella	0	0	0	0	0	0	0	0	0	0	0
Lime	0	0.2	0	0	0.1	0	0	0.1	0	0	
Thin-crust	0	0	0	0	0.2	0	0	0	0	0	
Fossilize	0	0	0	0	0.1	0	0	0	0	0	
Algae Colony	0	0	0.5	0	1.4	0	0	0	0	0	
Marginal Frequency Measure of Groups	2.6	0.7	5.8	1.1	4.2	0.1	1.8	0.9	1.1	0.5	

even though the sentences themselves dealt with topics belonging to other groups. A more thorough and unbiassed collection of data would have most probably reduced the number of inappropriate choices from ten to about two. Even as it was, out of the ten inappropriate choices, only eight were completely unsatisfactory, and the overall accuracy of the technique could be taken as 90% of the multiple meaning words in the test sample.

Concluding Remarks

The figure of merit technique has several advantageous features. It can be programmed very easily for use by a computer. It was found to be effective in the elimination of superfluous target equivalents in the test case of 21 sentences. While it is realized that this was a small sample, nevertheless the trend of the results indicates that the method will be equally effective with larger test samples. The effectiveness can be improved by collecting the data from a much larger sample than the one that was used in the above calculations. Such a collection of data could be done by means of a computer. By using automatic collection techniques, it would be possible to increase the number of words for which probability measures could be calculated, and at the same time make the data much more reliable.

The figure of merit technique was specifically developed for use with scientific articles. As such, it has only minimal application to non-scientific articles.

Even though the examples given above were translations of Russian sentences, the method as well as the data on probability of word occurrences can be used in the translation of material from any other language into English; or, by collecting necessary data, from any one language into any other language.

The most important principle on which the method was developed was the consideration of the probability of word occurrences in different scientific fields. This was a logical and fruitful approach to take in solving the problem of non-grammatical ambiguity in automatic language translation. It is doubtful whether a deterministic method can be developed to deal successfully with the multiple meaning problem.