[*International Conference on Machine Translation of Languages and Applied Language Analysis*,
National Physical Laboratory, Teddington, UK, 5-8 September 1961]

MULTIPLE MEANING IN MACHINE TRANSLATION.

by

AMELIA JANIOTIS and HARRY H. JOSSELSON

(Wayne State University, Michigan, U.S.A.)

ONE of the most crucial problems in machine translation of languages is multiple meaning. The present article Investigates the multiple meaning problems involved in machine translation from Russian to English in the area of mathematics as well as suggests means for their solution by appropriate computer procedures. The material discussed herein has been analyzed in the course of research supported by the Information Systems Branch of the Office of Naval Research and has been dealt with previous to this publication in the Second Annual Report (August 1960) of the Wayne State University Machine Translation Group. The present study enlarges on the previous discussion.

WITH the exception of certain ambiguities which are irresolvable in the source language[1], any concept in a source language may be expressed without ambiguity in a target language, even though it may have to be rephrased because of differences in linguistic structure[2] and/or

[1] For example, consider the two Russian sentences:
Профессор Иванова нашла решение этой задачи.
Решение её было одобрено Академией наук.
It is ambiguous, in the Russian, whether её refers to the woman or to the problem, i.e., whether the translation should be "her solution" or "its solution".
Another example comes from "Господа Головлёвы " by Салтыков-Щед
"Поверили его надзору подьячего", could be either "They entrusted him to the care of the clerk", or "They entrusted the clerk to his care."

[2] The sentence Его не было дома is rendered in English "He was not (at) home", not "Of him it was not at home". Likewise Корабль бурей разбил is translated "The ship was broken by the storm".

differences of linguistic perception[3]. A human being, versed in both languages as well as the subject matter being translated, is able to perform such a transformation of concepts. However, man is an exceedingly complex mechanism. In the case that a source language form has more than one target language equivalent, he is generally able to choose the appropriate equivalent(s), sometimes without even becoming aware of the irrelevant alternatives, on the basis of his orientation in the context. If, in the source language, it is necessary to consider a group of words, rather than individual words in order to extract meaning, a human being can do this automatically, even if the words of the group are not contiguous.

Because it is not known exactly how man is "programmed" to recognize meaning, in many instances it seems impossible to determine mechanically what factors in the environment of an ambiguity contribute unambiguously to its resolution. It is difficult to generalize mechanically about the proper choice of meaning for many individual source language forms, let alone generalize about classes of such forms. But this difficulty must ***be faced and handled*** in order to achieve a translation which is better than a simple list of all the alternatives for translating a sequence of forms.

The only mechanical generalizations about meaning will emerge from a consideration of the ***context*** of the form (or set of forms) in question. One kind of context may be called situational. The individual who receives the telegram "SHIP SAILS TOMORROW" will react differently if he is a manufacturer of equipment for small craft than if he is about to embark on a journey. (If he is a manufacturer about to embark, further consideration of his immediate affairs would be necessary). The recipient of the wire "ARRIVING TOMORROW WITH CHAOS" might be bewildered if he were not aware of the fact that his colleague was bringing to the West Coast a Chinese linguist (named Chao) and his family, Dr. Chao having been hired by the University of California at Berkeley.

The situational context cannot be used mechanically, since only the written text is available to the computer. Even if the situation is described in the text, it is a very complicated matter to search out the significant elements in previous sentences or paragraphs. However, the field of discourse, which may be thought of as a level of context, is significant. For example, the English word "pig" would have one

[3] In English we have a single word, "blind", to express incapability of seeing, but we do not have a single word to express the capability of seeing ("seer" has come to have a very specialized meaning). Hence, Слепец после операции снова стал зрячим. , must be rendered as "The blind man after the operation was able to see".
In the Russian sentence Льёт, как из ведра. , the notion of "buckets" has, as its English counterpart, "cats and dogs".

translation in a target language if it appeared in an article on animal husbandry, another if it appeared in a discussion about metallurgy (especially in the combination "pig iron"), and a third in a sociological treatise or in the literature dealing with American society of the Twenties (in the expression "blind pig"). (One might even regard "blind pig" as an idiom in the last-mentioned field, although certainly not in the first).

Besides the field of discourse, one must also consider the immediate context of a word. In a mathematical article dealing with partial differential equations, the Russian word степень may be translated by "degree" or by "power", depending on the immediate environment, similarly, величина may be "magnitude" or "quantity"

All of those contextual considerations are probabilistic; it is conceivable that in a novel about the Twenties there could be an incident in which, as a result of a drunken party, a few young men and women go out into the country, invade a farm, capture a squealing pig (who happens to be blind), hold him down, and attempt to feed this ***blind pig whiskey***. Even if one recognized the possibility of a blind pig being the animal in a Twenties' novel, and set up the word "whiskey" as one of the contextual criteria for deciding when it was ***not*** the animal, the test would fail, in this instance, to yield the proper translation[4].

The above failure notwithstanding, one must formulate rules like: Source language word X has target language translation Y_1 when any one of conditions C_{11}, C_{12}......C_{1n} holds (and the conditions are to be tested in the order listed). A condition C_{ij} may be a statement like: The source language word X_s (or a member of a certain class of source language words X_s) must be found after the word X with the possible intervention only of a member of the word class W5[5]. It must be remembered, also, that these

[4] In Russian, нос can mean "nose" or "cape" (in the geographical sense), and губа can mean "lip" or "inlet" (likewise geographically).

[5] Here are some simple examples of resolution of multiple meaning of items found in a text on partial differential equations:

часть - "side" when immediately preceded by левая or правая,
"part" in all other cases.

Следует - "it is necessary" when followed by an infinitive,
"it follows" in all other cases

означать - "to mean" when immediately followed by что ,
"to denote" in all other cases

из "of" when it occurs with состоять, строить,
образованный, составленный, второй,
каждый, максимальный, один,
первый, теорема, функция
"from" in all other cases

rules apply with a certain probability in certain narrow situations. The principal problem in connection with ambiguity is to work out practical schemes of syntactic analysis and semantic word association which keep producing better approximations to a (non-existent) ideal solution of ambiguity in translation. The semantic schemes will probably have to be worked out on the basis of a limited subject matter, even though a certain amount of general language is to be found in almost any scientific article.

The type of multiple meaning which has been discussed so far (which we may call "true" multiple meaning) is only one of several ambiguity problems which may be distinguished in working with a corpus. Ambiguity itself may be defined as a situation in which a form in the source language has more than one corresponding form in the target language in different occurrences, or a situation in which it is not sufficient to consider individual forms, but rather combinations of forms in order to achieve a meaningful translation. We have discovered[6] and classified the following types of ambiguity: (1) homographs, (2) inflectional ambiguities (applying to nominals, modifiers), (3) predicate block structure translations, (4) lexical idioms, (5) orthographic coincidences, and (6) "true" multiple meaning. These types apply specifically to the Russian —> English transformation. Any other pair of languages may yield different problems.

The first three types of ambiguity may be thought of as deterministic, i.e., the resolution of the ambiguity is accomplished or determined by examining the syntactic structure of the context. The last three types are probabilistic, which means that the ambiguity is resolved by examining combinations of words whose juxtaposition indicates, with a certain probability, that one of the words is to be translated by a certain one of various possible choices, or the entire combination is to be translated as a unit. Not all of the six classifications are mutually exclusive; for example, it is possible to resolve a homograph as functioning, in a given context, as a specific part of speech, and then find a "true" multiple meaning problem or a predicate block structure situation[7] within that part of speech.

[6] We used two articles for our study:

В.М. Борок. Решение задачи Коши для некоторых типов систем линейных уравнений в частных производных. Математический сборник 1955, Т. 36 (78) No. 2 and

И.М. Гельфанд и Г.Е. Шилов. Преобразования фурье быстро растущих функций и вопросы единственности решения задачи Коши. Успехи математических наук, Т. VIII выпуск 6.

[7] If надо is resolved as a preposition, then the problem of translating it as "over" or "above" remains. If аналогично is resolved as a predicative, there is still the problem of whether to translate it as "it is analogous" or "is analogous", for example, depending on whether or not a subject is present.

Homographs are identical *forms* which cross word class boundaries. They may be semantically related and predictable (as with аналогично which may function as a predicative, an adverb, or a preposition, and кругом which may be an instrumental nominal, an adverb, or a preposition), or they may be accidental (as with надо which is a predicative or a form of the preposition над, and дам which is a predicative (first person singular of дать) or a nominal (genitive plural of дама)). Their resolution is primarily syntactic — it may consist of an interrogation (a sequence of questions specifically related to the type of homograph, i.e., which word classes are involved) of the environment of the ambiguous form — an interrogation which, it is hoped, will lead to the proper choice by ruling out the other possibilities. Again, it must be remembered that it is not always possible to accomplish this mechanically.

Inflectional ambiguities are nominals or modifiers whose case, number, and for modifiers sometimes even gender, are not distinct. The ambiguity of a modifier may be reduced by comparing it with other modifiers modifying the same nominal, and with the nominal itself. Likewise, the ambiguity of a nominal may be reduced by comparing it with its modifiers. The ambiguity of a nominal block (a nominal plus its modifiers, including dependent adverbs) may be reduced if it is the object of a preposition, or if it can be associated with any governing structure. The Russian word станции may be genitive, dative, or locative singular, or nominative or accusative plural. In the sentence Машинист вводит поезда в станции., ("The engineer drives trains into (the) stations"), the agreement code of the nominal станции may be compared with the government code of the preposition в, and the ambiguity of the former will be reduced to locative singular or accusative plural, since the latter can govern only those two cases. Further, if the preposition is identified with the verb вводит, which is a verb of motion, we conclude that в governs the accusative in this case, and thus, that станции is in the accusative plural.

A predicate block structure is a form or set of forms functioning as a predicate. It can be classified structurally as simple or compound. A simple predicate structure consists of a finite verb (имеем), a short form predicative (легко) or phrase ([было , будет] легко), or an auxiliary [может] or phrase [можно (было, будет)] occurring without an infinitive complement. A compound predicate structure consists of a simple predicative plus an infinitive or infinitive phrase (будем иметь, может быть сделано, хочу видеть), or a simple predicative phrase plus an infinitive (можно будет сделат) or a string of infinitives (можно было продолжать учиться).

This composition may be schematized as follows:

I SIMPLE PREDICATIVES

a) finite verb forms: ALL HAVE INFINITIVES

читаем[8], хочешь[9], будет[8]

b) short form predicatives: HAVE INFINITIVES WITH БЫТЬ IN PHRASE FORM

1. verbal (short form past passive participles):

сделано[8], решено[9]

2. non-verbal (short form modifiers):

хорошо[9], аналогично[8]

c) modal and temporal auxiliaries: HAVE NO INFINITIVES

1. finite verb type:

может[9], можете[9], буду[10], будет[10]

2. short form type: HAS PHRASE FORM

можно[9], надо[9], нельзя[9]

II COMPOUND PREDICATIVES

Any of I (a[9] , b[9] , c[9,10]) plus infinitive or infinitive phrase, or any of II, where the last infinitive may take an infinitive complement, plus as many infinitives as style will allow.

Examples:

Ia[9] : Ты хочешь видеть.
Ib[9] : Было решено продолжать разговор.
Хорошо жить там.
Ic[9] : Я могу делать всё, что может быть сделано.
Этого нельзя сказать.
Ic[10]: Он будет это делать.
II : Можно было продолжать учиться.
Они хотят попробовать продолжать учиться.

This scheme of classification constitutes a basis for a mechanical scheme to identify and "block" predicate structures. It will also facilitate identification of the patterns of combination which are in constant, but non-parallel relation to the corresponding English patterns. For example, the form сделано is translated as "done" when it occurs with an

[8] Does not take infinitive complement

[9] May take infinitive complement.

[10] Serves as a temporal auxiliary, and as such takes imperfective infinitive complement.

auxiliary, "is done" when it occurs with no subject and no auxiliary, and "be done" when it occurs with пусть . A Russian infinitive will be translated with the English word "to", except when it occurs with one of the auxiliaries.

An ***idiom (lexical idiom)*** may be defined as a structure whose translation is not equal to the sum of the translations of its elements taken separately. As examples, consider the expression в конце концов which would be literally rendered as "in (the) end (of) ends" but idiomatically as "finally", or подинтегральная функция which could be "subintegral function" meaning "function under the integral (sign)", but is much more elegantly translated as "integrand". As a subclass of the lexical idioms one may recognize the government of prepositions by predicatives or verbal derivatives. We had the following examples in our text[11]:

A)	нарушаться от	(to) be disturbed by
B)	обращаться в	(to) become
C)	сократить на	(to) cancel by
D)	ССЫЛАТЬСЯ НА	(to) cite
E)	убеждаться в	(to) be convinced of
F)	указать на	(to) cite

(In addition, we had an instance of multiple meaning within a lexical idiom: the forms входит в and входить в had translations "belongs to" and "(to) belong to" respectively, the form входят в had translations "belong to" and "enter into", and the forms входящие в and входящих в had translation "occurring in".)

B, D, and F are instances in which the translation of the preposition is suppressed, whereas in A, the basic meaning of the preposition, "from", is replaced by "by", in C, the basic meaning, "on", is replaced by "by", and in E, the basic "in" is replaced by "of". B might have been translated as "(to) turn into", and F as "(to) point to"; in both cases, then, the basic meaning of the verb would have been retained, and in B, the preposition would have had its basic meaning "into" which holds in the presence of a verb of motion, whereas in F, the preposition would have to assume a secondary meaning, "to", rather than "on".

Orthographic coincidences occur when two unrelated stems generate forms which are orthographically identical, but which, unlike homographs, both fall within the same form class, or when a single stem generates orthographically identical but morphologically or semantically distinct forms.

[11] И.М. Гельфанд и Г.Е. Шилов. Преобразования Фурье быстро растущих функций и вопросы единственности решения задачи Коши. Успехи математических наук, T.VIII, выпуск 6.

Examples of the first type are: пла́чу, "I weep", from пла́кать, and плачу́ "I pay", from платить: мука́ = "flour" and му́ка = "torment, the latter from the verb му́чить: зреть, "to behold", conjugated зрю, зрит, ... and зреть, "to ripen", conjugated зрею, зреет, ...; вы'купать, "to bathe", perfective of купать and выкупа́ть, "to redeem", perfective of выкупить. Examples of the second type are: отреза́ть and отре́зать , "to cut off", imperfective and perfective respectively; сбега́ть, the imperfective form of сбежать and сбе́гать, a perfective form which has no imperfective — the first means "to run down (stairs, a hill)" while the second means "to run (and deliver or fetch something)" and is used with в or за.

In our treatment of ***"true" multiple meaning*** in the text, we found that we were unable to solve many of the instances, and that we were able to find only partial solutions in other instances. These cases will require human postediting on output where all possible meanings will have been listed, until such time as human ingenuity invents mechanical schemes of great complexity, first for individual word solutions and then, if possible, for word class solutions of multiple meaning. Определить and its derivatives are, at present, impossible to resolve because of the difficulty in distinguishing "definitions" from "determinations (of value)"; in some cases the solution would necessitate an examination of a mathematical formula. Изменение presents a problem because English distinguishes between "change" and "variation"; we speak, of "***changing*** the order of integration" and we say "without ***changes i***n the proofs", but we have a "***varying*** argument", or a "domain of ***variation***", and we say that something "***varies*** within the limits", while Russian uses the same semantic unit for both of the above concepts.

In the preceding discussion we have already given indications of how the six types of ambiguity may be resolved. We will discuss our approach to their resolution further in terms of computer procedures.

The ***homograph resolution*** pass takes place prior to the main syntax programme. It will be a set of syntactic subroutines — one for each type of homograph. It may duplicate some of the later syntactic analysis, but it is wise to keep it separate, at least in the beginning, so that the homograph phenomena may be studied in isolation.

The ***inflectional ambiguities,*** on the other hand, may be reduced during the syntax programme in such parts as "nominal blocking" and "prepositional blocking" as described above in the paragraph dealing with inflectional ambiguity, and ultimately solved by further questioning (as, for example, relating a preposition to a verb of motion, or finding an unambiguous nominative which may lead to the resolution as accusative of a nominative-accusative ambiguity on the other side of the predicative).

The ***predicate block structure*** can be determined by a special routine designed for that purpose, which will probably follow the nominal, prepositional, governing modifier blocking cycle which we propose to execute in our programme.

The ***lexical idioms*** should be stored ideally in the dictionary (it will be necessary to lookup the longest form of any given sequence in text, i.e. if имеет место occurs, and this sequence is in the dictionary, the look up will not stop with имеет but rather go on to find the whole idiom) and looked up as they appear in text. It will subsequently be necessary to devise a scheme for recognizing non-contiguous idioms.

Orthographic coincidences must be treated in the same way as ***"true" multiple meaning*** — the minimal amount of meanings which will cover all empirical circumstances must be chosen, and further reduced, if possible, by consideration of the subject matter (micro-vocabularies). If ambiguity remains, the environment should be searched for markers, both semantic and syntactic, which can be set up experimentally and periodically improved and refined.

It is abundantly clear, therefore, from the foregoing that in order to handle the resolvable types of multiple meaning occurring in the course of machine translation from one language into another it is necessary to classify and list all the occurring types, both probabilistic and deterministic, and then laboriously, step by step, develop the proper algorithms and computer routines for dealing with them. That these will vary with each language pair involved goes without saying. Only by these procedures will one be able to handle this all-important problem of multiple meaning in machine translation of languages within the limits delineated above.