Session 1: CURRENT R ESEARCH

RESEARCH IN MACHINE TRANSLATION AT RAMO-WOOLDRIDGE [1]

Jules Mersel
Ramo-Wooldridge Laboratories

Purposes of the Research

The primary purpose of the work that we are doing is the development and use of techniques of systematic language research. We are committed to the belief that by making optimal use of computing machinery we can make significant progress in areas of linguistic research where heretofore progress has been severely hampered by practical decisions.

In particular we are committed to the exploration of machine translation research techniques and are placing emphasis on semantic problems. Among our goals is the development of a technique for practical automatic translation. It is not necessarily our purpose to try to create the high quality of translation that one has a reason to expect from a human translator who is not only adept in both languages but who also has a great familiarity with the subject matter he is translating. We, of course, have no quarrel with lofty purposes per se, but we believe we can get useful results much sooner. Despite the many examples of human translation of this quality, we are seriously skeptical about the maintenance of this standard as the demand for a large amount of up-to-date translation mounts. Consequently we are satisfied by trying to attain translations which will be more useful to the reader than no translation or a translation that he receives months too late.

Within the scope of our studies in machine translation, it is our purpose to conduct basic research in the area of meaning in language. In particular we are committed to research in how to ease the problem of multiple meaning when one translates from Russian to English.

[1] The work in machine translation at Ramo-Wooldridge is sponsored in part by the Rome Air Development Center of the U. S. Air Force, and in part by the Ramo-Wooldridge Research Program for 1959 and 1960.

Research Techniques

It is not our belief that one attempts to solve all problems before going to the computer. We do not even believe that the seriousness and hence the priority of many problems can be ascertained without the aid of a computer.

It is our procedure to start by using the expert knowledge of both Russian and English that members of our research team have in order to gain initial understanding of the problem and its probable solution. We then mechanize the solution and test it out on our computer. We never expect that we will succeed in completely erasing a problem with this try. What we are interested in doing is getting a verification of how well our solution worked and in exposing the next most frequent problem area.

It is in our devotion to minimize, rather than to exercise problems, that we find our greatest philosophic differences in our discussions with linguists primarily devoted to academic work. We seek neither a complete solution nor a solution of those problems of greatest linguistic interest. We seek instead to cut to low frequency those mistranslations, ambiguities, or unidiomatic results which plague us the most.

If a problem area appears every 5, 000 words, its solution does not concern us unless the solution is both obvious and easy to apply. However, a problem that appears every 100 words does interest us despite the difficulty that may be associated with applying the solution.

We do not believe that the mere running of large amounts of text will by itself solve any problems in translation. Problems are still solved by people and not by machines. We do believe, however, that the running of large amounts of text serves to test previous solutions, allows one to get a better feel for the most frequent problem areas (i.e. , when seemingly individual problems form a cohesive and easily attacked unit), and gives sufficient examples of words in actual usage to allow an attack on their multiple meanings. Furthermore, the speculative powers of linguists and informants are limited, and the ability to collect data by machine extends their powers considerably.

Our research technique is of a cyclic nature. It is a cycle of observation, idea, mechanization, test, correction and new observation. In this cycle there is a use of both humans and machines. In

the beginning of our work the actual machine involvement was small. As we have progressed, we are making headway with our predetermined plans to mechanize more and more of this cycle. In our previous report we have frequently shown a block diagram of our research cycle. The mechanized blocks are shown in red and the blocks denoting human activity in black. We are gradually changing more and more blocks from black to red. We thus attempt increasingly to free our people from drudgery so that they may devote themselves to creative work. A decrease in drudgery from 95% to 60% may still leave a lot of drudgery, but it gives a very pleasing increase, by a factor of eight, in the amount of time available for creative work.

Problem Areas

I would like to list those problem areas of today's machine translation that we are attacking. The list will not be in any order of importance, but will be in the order that they appear in a machine translation program.

INPUT: One of the formidable obstacles to practical machine translation lies in the area of transcribing the original source-language text onto a machine-readable medium. Today's solution of either keypunching or unityping is unsatisfactory both from the point of view of economics and from the point of view that diminishing drudgery is a good thing. In the beginning of 1959 keypunching cost us 3¢ per word. By mechanizing the pre-editing and by replacing keypunch verification by sight verification, we have reduced the cost to 1.1¢ per word. When you consider that 1.1¢ per word is what a professional translator gets from a translating service, one gets a feel for how economically unfeasible a solution to the problem is presented by keypunching.

Other R-W research and development is addressed to the problem of the automatic transcription of text.

DICTIONARY LOOKUP: From the point of view of ability to solve the problem, the dictionary lookup of the text words presents an easily solved problem. The problems here are not those of feasibility but those of economics. It has always been possible in theory to keep a full-form dictionary. Those of us who are facing the problem today are concerned with linguistic techniques that will keep down the size of the dictionary and programming techniques that will

keep down the time of the search.

In attempting to keep down the size of the dictionary we have not gone all the way to a single-form-per-stem glossary. Our glossary averages around two and one-half forms per stem. When a form comes in that is not in our glossary we attempt to find its stem and we seek a matching stem in our glossary. If successful, we use the grammar code of the stem in the glossary and an algorithm based upon the ending of the text word to create a grammar code for the next word. Our procedure has been highly satisfactory. In the last group of 10, 000 words of text that we translated we had only one occurrence of an incorrect code assignment.

Of course, if a stem is missing from the glossary, it is necessary to add it to our tape with all its surrounding information. We do not attempt to detect such words before a run. It is important to us to determine just how well we can handle the rest of the sentence despite this missing or misspelled word. We have been reasonably successful in the fabrication of grammar codes for missing words. We have created a routine which, on the basis of the ending of the missing word and on a basis of the examination of its four surrounding words, supplies a provisional grammar code.

This routine was tested on a text of 9,641 words. Each word was treated in succession as a missing word. A total of 6,758 words had a grammar code assigned which would have been adequate for the proper functioning of our syntax-analysis routine. Considering the large number of conjunctions and prepositions in the text, and the fact that the routine was not designed to take care of these few but frequently appearing words, we were quite pleased with the result.

I have indicated that a word was not added to the glossary until after it had shown itself missing in a translation from text. The question then arises as to which form should be added. We take an interesting stand in this matter. We add all forms that appeared in the text. Since fabricating a grammar code is more expensive than just finding a word, we desire to supply our dictionary not with some canonical form but with the forms most likely to appear in text. Once, however, that a stem makes an appearance in the glossary, there are no additions of forms if the grammar codes for those forms can be supplied by our usual procedure.

The programming techniques to effect the dictionary search

depend on the size and speed of the internal memory that is available. Tomorrow's enormous high-speed memories will certainly change our attack. We have seen this already as we have advanced from an 8,000 word 704 to a 32,000 word 709. Here we are getting an increase of a factor of eight in lookup speed while only paying an increase of 40% in computer rental. This increase is partially due to a larger memory and partially due to the ability to read and write tapes while carrying out other computations. The memory itself has received no increase in access time. We expect another factor-of-four benefit from putting our program on a 7090. Here there will be no increase in memory size, but there will be a four-to-five-fold improvement in access time from the tapes and cores. During the period of two years we will have seen a 32-fold increase in lookup rate with only a doubling of computer cost. The cost of the dictionary lookup will no longer present an economic argument against the practicality of machine translation.

IDIOMS: The question as to what to classify as an idiom keeps coming up again and again in our work. The difficulty is that this kind of classification provides such a conceptually easy way out of many difficulties. In the extreme, even complete sentences could be considered as idioms and hence a difficult syntactic or semantic problem could be seemingly dispensed with. Possibly because we have been repelled by the search cost, we have tried to keep our list of idioms small. If the meaning is clear and the occurrence is not frequent, we do not add to the list of idioms in our computer. The question as to whether to translate a certain phrase literally as "... from the editorial office of" or idiomatically as "... edited by. . . " would be a case decided by the frequency of the occurrence.

When we do add an idiom to our list, we go further than merely supplying the idiom with an idiomatic English translation; if the individual grammar codes of the words that constitute the idiom do not represent the grammatical import of the idiom, we supply the idiom with its own grammar code.

As yet we have made no attempt to handle those idioms whose component words are separated by other text words. We have restricted ourselves to idioms whose words appear in an unbroken string.

Among our idioms there is a large class whose handling is

both inexpensive and easy. These are the two-word idioms whose correct translation requires only an inversion of word order. Though they are detected during the idiom lookup, their translation requires only that a flag be supplied to the word-order routine.

SYNTAX AND MORPHOLOGY: Our differentiation between syntax and morphology is an operational one. For our purposes, the morphology is the grammar code supplied by the glossary, the missing-form routine, or the idiom routine. All the rest is contained in our syntactical pass.

Our syntax analysis and its resulting translation decisions are heavily influenced by the work of Professor Paul L. Garvin. For those who are interested in the philosophy of this analysis, I refer them to his talk in Session 6 of this Symposium.

Our syntax routine received its original form as a syntactical discriminator. Its function was to separate the syntactically simple sentences from those which presented syntactical difficulty. It soon became obvious that with modification this routine could provide a useful syntax routine for our machine translation program.

Though for purposes of our missing-form routine, we rely heavily upon the traditional parts of speech classification, for purposes of our syntax we use a different grouping. Participles, adjectives, and certain pronouns are all grouped together as modifiers. Nouns and certain other pronouns are grouped as nominals. Infinitives and gerunds each form their own class. The other forms of the verbs are grouped with the short forms of the adjectives and participles as predicatives. Relative pronouns form a class of their own.

The first thing we do in our syntactic pass is to examine the neighborhood of the symbols and numerals in text. The result of this examination is an assignment of a grammar code to these entities.

The second thing we do is hunt for homographs. Though the fact that a word is a homograph is indicated in our grammar code, its resolution is made during the syntactical pass. Our definition as to what is a homograph depends upon our definition of the parts of speech.

The third phase of the routine is a search for and labeling of inserted structures.

The fourth pass is a search for and a labeling of governing-modifier packages.

The fifth and last preliminary pass is a search for relative-pronoun packages.

Having finished all these preliminary passes, we start our major syntax analysis. Our major syntax analysis depends upon finding a predicative or gerund to use as the pivot of the sentence. Having found this pivot, we then hunt for subject and object packages. In the course of this hunt, other packages such as prepositional packages are found and labeled. Great use is made of the government characteristics of the predicative. Appropriate translation decisions are made throughout the syntax routine.

SEMANTICS AND MULTIPLE MEANING: The problem of meaning and the resultant choice among the possible English equivalents of a Russian word provides one of the most interesting and important subjects of research for the field of machine translation today. The problem is further complicated by the fact that writers do not always say what they mean.

In general, our translation program provides data to multiple-equivalent research so that the research may later provide rules for our program.

Two significant aspects of our research on multiple meaning are appropriate to a brief summary type of statement. At this stage of our work, a major part of our effort has gone into direct examination of numerous specific problems of multiple meaning. Based upon this examination has been the development of formats for data collection and the design and implementation of a research procedure for semantic studies. Thus, we follow a largely inductive approach wherein we attempt to develop from the collected data general patterns representative of multiple-meaning problems. At the same time, however, and in parallel with this effort, we have hypothesized several models which form the basis for a more deductive approach.

A satisfactory discussion of this work is not possible in the time remaining for this talk. A description of our work in this area will be given by Don Swanson in his talk in Session 9 of this Symposium.

RESEARCH TOOLS: It was my thesis earlier in this paper that

the primary purpose of our work in machine translation was not simply the production of a machine translation program, but mainly to do linguistic research, primarily with emphasis on semantics, by use of computing machinery.

In this section of my talk, I would like to explain the major tool that we receive from our translation program.

Though our routines provide very detailed listings of the results of our translation program, these listings are not in a convenient form for future manipulation. Consequently, we decided early in the design of the program that, in addition to the listings, we would record onto magnetic tape a complete history of the translation and would then update this history with additional information obtained by postediting the translation. This tape is called the information tape and is to be the prime source of most of our listings.

The information tape contains a full corpus of text. The total number of words for all the articles in this corpus could be as high as 25,000 words. The tape is divided into variable length records. Each record represents a sentence of text. Within the record each text word is represented by 19 computer words.

At this point, I call your attention to the 19 computer-word items.

Words 1-14 contain, with some additions and changes, the original glossary item.

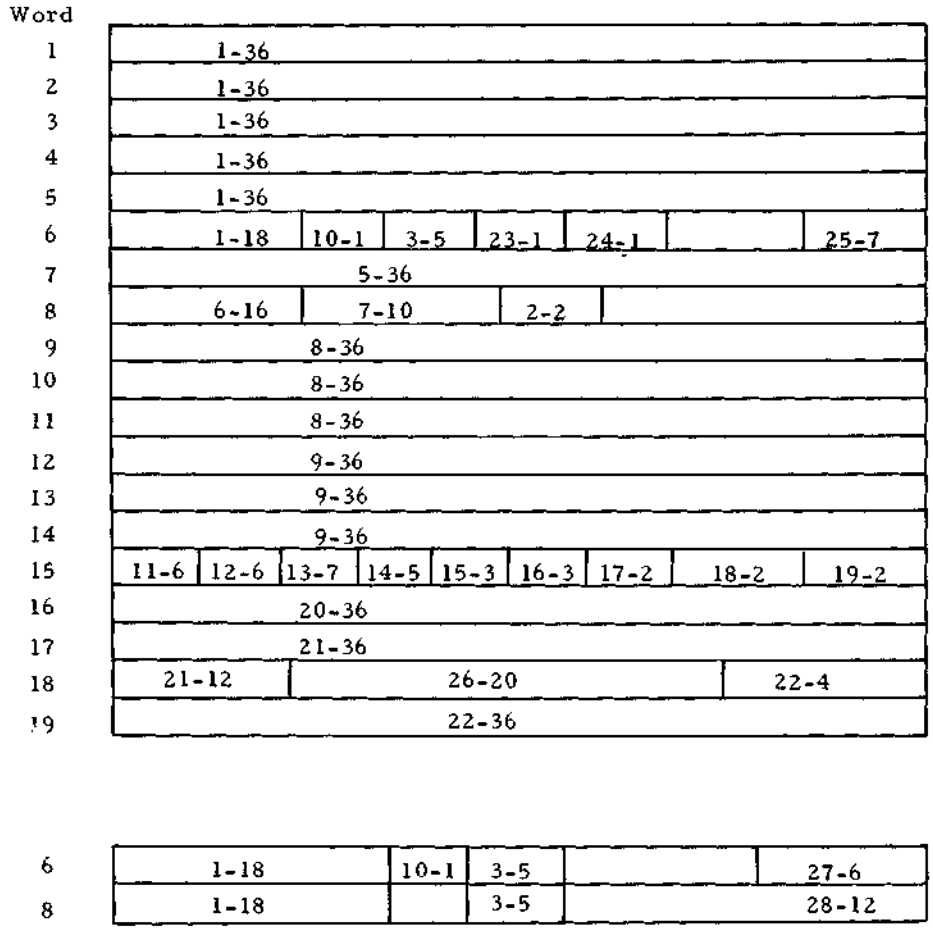
Word 15 is the result of our pre-editing pass. It contains text location and punctuation information.

Words 16-19 are created during our translation and they indicate syntactic, semantic, and translation decisions.

Field 1, which stretches from word 1 to the middle of word 6 (the number after the hyphen stands for the number of binary digits used in that word), contains originally the possible English equivalents of the Russian word. If the program makes a decision as to which equivalent to choose, the rejected equivalents are erased from this field. The field contains room for 33 characters, and hence is more than sufficient for the result. It can become a bit crowded, however, in its initial glossary form.

Field 1 can also become modified as a result of decisions by the posteditor. When the posteditor, as a result of his examination of the translation, decides that certain equivalents are an

Figure 1 - COMPUTER WORDS



incorrect translation for that portion of text, only those equivalents which the posteditor feels are acceptable are allowed to stay on the tape. A word is not rejected because of the posteditor's sense of style. A rejection indicates complete unacceptability.

Since the posteditor rejects not on a basis of style, it is possible to expect him to indicate what is in his opinion the determiner for his choice of the correct equivalent. We have created a convenient notation for this purpose. The notation can indicate that the determiner was the second noun towards the beginning of the sentence or the following verb or some other form of combined syntactic and positional information. Of course all his reasons might not be so easy to represent. As a result, one notation indicates that the reason for the choice is not written onto the edited text but is entered into a separate posteditor's log.

This procedure of not making the posteditor make a choice from among acceptable equivalents has many advantages. It reduces the difficulty the posteditor has in indicating the reason for his choice. It also might save the analysis from misleading statistics.

For instance, in a two-equivalent word, on a basis of style, the posteditor might have chosen the first equivalent 60% of the time and the second equivalent 40% of the time. On the basis of mandatoriness, however, the first equivalent might have been dictated only 3% of the time while the second equivalent was demanded 10% of the time. Thus could be changed the basis for an eventual semantic rule.

Field 10, a one-bit field, indicates whether the grammar code for the word was created by our stem-affixing procedure.

Field 3 contains five bits and indicates the length of the stem. This is computer-determined in our glossary-maintenance routine by the inverse of the stem-affixing procedure. This information is used in our missing-form routine. The information could have been derived there, but in the interest of speeding up the glossary search the information is placed in the glossary once and forever.

Field 23 contains one bit whose function is to indicate whether that word participated in an idiom in that sentence.

Field 24's one bit serves the function of indicating that the grammar code is not the same as the original glossary entry. This could be the result of stem-affixing, our missing-word routine, idiom participation, or agreement checks made during the syntax routine.

Field 25 indicates the correct English word order. This variance from the Russian order might have been decided upon either in the idiom routine or in the syntax routine. The Russian word order is implied by the position of the 19 word item within its record.

The last 12 bits of word 6 serve different functions earlier in the routine. In their original glossary form they contain information for the missing form routine. Later on they contain relative text location information which allows the routine to get the text back into text order and out of alphabetic order.

Field 5 contains 36 bits whose function is to indicate multiple-choice rules. This field is far from filled. In the future we expect this field will need to grow beyond its present limits. As will be explained we have left room for growth.

Field 6 contains the stem number of the word. It is the same for all forms of the word. Its 16 bits allow for an eventual 64,000 word stem glossary before overhaul is necessary.

Field 7 indicates the form number of the stem. It is also created by the inverse of the stem-affixing routine. Its unnecessarily large size could allow for the expansion of field 6. Its primary use to date is to allow the distinction between forms for the idiom routine.

Field 2 has the purpose of indicating whether a word in general is capable of being in one of our idioms. It is to be distinguished from field 23 which indicated whether the particular occurrence was idiomatic.

Field 8 contains the grammar code. Our code is spread out into bit form for easy Boolean Algebra manipulation during the syntax routine. When we get more information about the total number of different grammar codes that we will encounter in the language, it should be possible to cut this field down to around 12 bits for storage purposes and to expand it by table lookup during the routine. Its reduction in size will leave ample growing room for field 5's semantic rules.

Field 9 contains the Russian word. Its size allows the processing of words that contain up to eighteen Cyrillic characters.

Field 11 contains the text-article designator; field 12, the page indicator; 13, the line number; and 14, the word number on the line. This information is of use to our posteditors.

Field 15 indicates the punctuation before the word, while fields

16 and 17 allow for two punctuation marks after the word. Field 18 indicates the capitalization of the word. Field 19 indicates whether the word started a paragraph, names of authors, or title of an article.

Field 20 contains six characters for insertions after the word such as "-ly".

Field 21 contains eight characters for insertions before the word such as articles, prepositions, and auxiliaries.

Field 26 allows 20 bits to show all the syntactic packages and subpackages that a word might have fallen into. Thus a particular word might have been included in a governing modifier package, prepositional phrase, and nominal block.

Field 22 allows for the indication of up to five syntactic and semantic decisions that were made on the basis of that word. We believe the inclusion of this field to be of the utmost importance. In the past it was possible for us to get into the position where we knew a rule had failed 15 times but did not know whether it had been applied only 15 times or 2000 times. Obviously two different courses of action would be indicated for the two cases.

Field 22 allows us to count the number of times a particular rule has been applied.

By manipulating our information tapes it will now be easy to make requests such as for a listing of the most frequent nouns when they are followed by a genitive phrase or by certain modifiers.

Our information tapes, though they will be permanently retained in their text order, will be periodically sorted into alphabetic order and merged with the concordance of all previously encountered words in our text. From this concordance, it should be possible for our researchers into multiple meaning to find all occurrences of a word in our text with a text reference, the posteditor's decision as to possible English equivalents, and an indication as to the reason for the decision.

Source of Text

Up to now our translation has been from articles on physics taken from Soviet journals and books. Also we have keypunched and are continuing to keypunch certain other articles on physics of a more general nature. It was felt that in our first attempt to translate from a non-technical source it would be best to restrict ourselves to a subject where we would not be hampered by too many words missing

from our glossary.

I regret that in the time allowed it has not been possible to explain how our translation program actually works. A general report giving a detailed description of our program is currently in preparation and will be available in mid-March of 1960.