SOVIET RESEARCH IN MACHINE TRANSLATION

Kenneth Harper
University of California, Los Angeles, and The RAND Corporation

A survey of the literature indicates that Soviet researchers are attacking the problem of machine translation on a broad front, with considerable numbers of well-trained specialists in linguistics, mathematics, and computer science. Substantial achievements have been recorded in the analysis of input text, in the development of specialized glossaries, and in the creation of analytic and synthetic translation programs. This work has been performed chiefly on an empirical basis. Considerable attention has also been devoted to theoretical questions, particularly in the area of an intermediary language. The application of machine techniques in this research appears to be limited. The adequacy of existing glossaries and algorithms for translation of new texts has not been demonstrated. A significant breakthrough in automatic language translation can be expected if, and when, computer facilities are made available.

This paper does not purport to trace the history and development of Soviet research in MT, nor to catalog the various activities and actors in this field. Most members of this Symposium are probably more or less familiar with the papers published by Soviet MT researchers. The volume of such publications is considerable; my purpose is to attempt a summary and, if I may, a critique, of the effort. The basis of this attempt is extensive reading and a smaller amount of personal contact with some of the more active workers. I may merely hope that the objectivity of the summary and critique will not be seriously colored by my personal opinions and predilections.

Discussion of the state-of-the-art, present and presumed, should be preceded by a few remarks on the factors of quantity and quality in the Soviet effort. Anyone who has followed the literature will be impressed by the growth or magnification, of MT research in the Soviet Union. In 1954 there were a handful of people experimenting with the first English-Russian and French-Russian algorithm [l, 2]. In 1959, the workers engaged in full- or part-time

research in the field could be numbered in the hundreds. Thus, the 1958 Conference on Machine Translation in Moscow was attended by 340 representatives of 79 different institutions, including 21 Institutes of the Academy of Sciences of the USSR, 8 Institutes of Union Republic Academies of Science, 11 universities, and 10 other higher schools [3]. Seventy papers were read and discussed at the Conference. In 1959, at a more restricted Conference on Mathematical Linguistics in Leningrad, a total of 61 papers were delivered by some 78 participants. The majority of these papers reflect the serious, active participation of the authors in MT research or in closely related fields [4].

To judge from the literature, the most intensive efforts appear now to be concentrated in four working groups; the Institute of Precise Mechanics (Moscow), the Electromodeling Laboratory of the Ail-Union Institute of Scientific and Technical Information (Moscow), the Steklov Institute of Mathematics (Moscow), and the Experimental Laboratory for Machine Translation (Leningrad). The Institute of Linguistics (Moscow) appears both to sponsor research on the problem, and to supply trained linguists to these and other groups. Smaller groups working on MT are located at the First Moscow State Pedagogical Institute of Foreign Languages, and at the Universities of Moscow, Gorky, Kharkov, Kiev, Petrozavodsk, Tiflis, and Erevan. Related areas of study are represented by researchers from a variety of institutes specializing in problems of speech, psychology, linguistics, communication, and computer technology.

This is an impressive list. In an interval of four years the Soviet Union has developed a corps of trained workers in the MT field substantially larger than that of all other countries combined. Considering the shortage of machines for data retrieval, one may also assume that large numbers of language-clerical people have been employed; the acquisition of experience and "point of view" by workers on this level will certainly be an important factor in future research.

The breadth of the assault on the MT problem, made possible by the size of the corps of workers, is well known. Whereas in 1955, 2 language pairs were being studied (English-Russian and French-Russian), in 1959 some 20 pairs were under investigation [3].

Microglossaries have been prepared for a number of technical fields, and in a number of languages. The presumed quality of this work will be discussed later; for the present, it should be emphasized that the Soviet effort is broad enough to permit multiple studies on single language pairs (with attendant advantages of diversity and disadvantage of duplication), and at the same time, studies of many more languages than other countries have been able to afford.

In addition to numbers, the Soviet MT effort is characterized by the high quality of researchers. The main working groups include such people of authority and prestige as Professors D.I. Panov, and A. A. Lyapunov (IPM) and faculty members of Leningrad State University (Prof. M.I. Steblin-Kamenskij, Philology; G. S. Tsejtin, Mathematics and Mechanics; V.P. Berkov, Philology; N. D. Andreev, Philology). Involved in one way or another are faculty members of Moscow State University (V. V. Ivanov, V. A. Uspenskij, R. L. Dobrushin, O.S. Akhmanova), and recognized scholars associated with the Institute of Linguistics, ANSSSR, and the First Moscow State Pedagogical Institute of Foreign Languages (P. S. Kuznetsov, A. A. Reformatskij, V.I. Girgor'ev, I.I. Revzin, V. Yu. Ruzentsvejg).

The influence of men such as these has been responsible for the formation of an Association for Machine Translation (1956) and the holding of two national conferences. High-level seminars have been held on the future of MT and its relation to other areas of research. As early as 1956, the Philological Faculty of Moscow State University sponsored a seminar on mathematical linguistics, under the direction of P. S. Kuznetsov; the ideas of Academician A. N. Kolmogorov and others were discussed, with respect to application of mathematical research methods to linguistics [5] . In April, 1958, another seminar was held by the Institute of Linguistics to consider Ivanov's paper, "Theoretical Linguistics and Applied Linguistics"[6]. Liaison was established in 1958 between the Association for Machine Translation and the Committee on Applied Linguistics, ANSSSR.

The point to be stressed here is that with respect to organization and leadership, MT has had the interest, backing, advice, and sometimes participation of a number of the most prominent scholars and scientists in the country. The result has not been a hamstringing of individual research initiative, since these men were as confused or divided as anyone else during the first, groping years of

investigation. The familiar pattern of scholars arising to deflate the enthusiasm of the first, rash proponents of MT has been repeated in the Soviet Union. Thus, L. S. Barkhudarov and G. V. Kolshanskij in early 1958 cautioned that machines will find it difficult to translate "red herring" and "French supply troubles in Syria [7]" The interesting thing is that these same scholars retain their interest in the field long enough to be convinced, and perhaps to make some contribution[4] . It takes time to condition scholars (particularly linguists) in this area, and it may be that the Soviets have passed through this period and now possess a number of highly qualified people who can intelligently and positively apply their sophistication to the sometimes naive assumptions of MT researchers. If so, the Soviet MT effort may be entering on a second phase, to be characterized by an increased amount of direction and concentration.

It should also be said that on the working level, Soviet MT groups are very well supplied with linguists and people knowledgeable in languages. The contribution of mathematicians will also soon be evinced; in this regard, the establishment of a faculty of Mathematical Linguistics at Leningrad University in 1958 is likely to be significant.

The activities and achievements in Soviet MT research may be summarized under two headings: glossary construction and linguistic analysis.

Glossary construction

A number of topical glossaries have been compiled by various working groups in the past five years. In some instances, these glossaries are a by-product of empirical analysis of texts for the building of translation algorithms; in other instances, they are independent lexicographical studies. Typical examples of existing glossaries are the following:

1. A French-Russian glossary of mathematics, compiled from a text of 20, 500 running words (O. S. Kulagina, Steklov Institute of Mathematics) [2]. After eliminating words which occurred less than four times in this text, and after adding words deemed to be essential, a stem glossary of 1, 200 items was attained. I have found no discussion of the adequacy of this glossary in tests against new texts. This is apparently the glossary mentioned by Panov in descriptions of the earliest French-Russian algorithm [3].

2. An English-Russian glossary (not a stem glossary) in the

field of applied mathematics, composed of some 2, 300 English items (Belskaya) [9]. A list of six English books is given as the text source. A text of this size would approximate a quarter of a million running words; unfortunately, Belskaya does not specify that the whole text was used in compiling the glossary, nor that all words encountered in the survey were retained. In view of the magnitude of a "hand" survey over a text of this expanse, and in view of the relatively small size of the glossary, we may presume that the treatment was "selective", both in text examination and in retention of items for the glossary.

3. A German-Russian glossary of unstated size, for which "some 1, 500 pages of mathematics were studied," (S. S. Belokrinitskaya) [9] . The statement is made that the glossary "comprises, without exception, every word which we took into consideration". This wording indicates that not all words encountered in text were recorded, or at least were not included in the glossary. (If this interpretation is in error, i.e., if all text occurrences are represented in the glossary, this is the largest microglossary compiled anywhere.) The glossary, and the text source, are described in terms of a study of multivalence in German-Russian translation.

4. An English glossary of geology of 7, 535 words (M. G. Udartseva) [10] compiled from text sources comprising 250,000 words. It is stated that the frequency of occurrence of each word is recorded, and that a "minimal glossary" of 2, 373 lexical items was formed. Of these, says the author, "176 are specialized terms; more than 200 words have another meaning in geological literature, while the remainder are ordinary words. About 4, 000 of the 7, 535 words are technical terms". When the minimal glossary was tested against random text, 1 to 1-1/2% of the text occurrences were unrecognized. Against texts in the area of political discourse the figure reported is 8 to 10%, and against Dickens, 16 to 18%. The minimal glossary was compared with Thorndyke's dictionary. The glossary was compiled by hand (there are references to index cards), and apparently has no immediate connection with a MT algorithm.

Other smaller glossaries are mentioned: Mel'chuk's Hungarian-Russian mathematics glossary, an English-Russian stem glossary mentioned by T. N. Moloshnaya (perhaps identical with No. 2, above), Chinese, German and Japanese mathematics glossaries (Belskaya, 9), etc.

The presumed characteristics of these glossaries are as follows:

1. They are text-based, i. e., not compilations from published dictionaries. The extent of foreign texts surveyed in four of the above-mentioned glossaries is approximately 800, 000 running words.

2. They are compiled by hand. The literature contains no indication that the running text has been recorded on punched cards or tape. A considerable expenditure of man-hours is implied.

3. A decided preference is shown for stem glossaries, and for "minimal" glossaries of between 2,000 and 3,000 entries. The main consideration here is apparently the limitations of storage in Soviet computers.

4. Accurate and detailed frequency studies have, with some exceptions, not been made. This is a consequence of the compilation method. It is not clear that glossary compilers have recorded precise sequence numbers for each occurrence in text, except for words which were deemed worthy of study; in any event, "hand" retrieval of context is quite laborious for a large volume of text. Context study has been limited to instances of homography, multivalence, etc. In general, the utility of this material for future study appears to be limited.

5. The adequacy of existing glossaries for new text is good, as one would expect. Belskaya reports "one or two" unrecognized words per page of new text (apparently less than 1% of the running words, although it is not clear that this figure includes repetitions of unrecognized words; it apparently does not include proper names). Udartseva reports 1 to 1-1/2% new words in texts from a broader field--geology. These figures are comparable to those obtained at The RAND Corporation when a new physics text in Russian was compared against a physics glossary of some 6,400 words. Other Soviet groups do not report figures. There is no mention of provisions for updating the glossary as new text is tested. Researchers are generally disposed to accept microglossaries against which all except one or two percent of occurrences in new text will be recognized. Indeed, Belskaya states that the latest phase of her work will be concerned with the building of microglossaries for a number of new fields. Such projects are entirely in keeping with the limited • tore available, and indicate the goals of MT in the near future: translations of subfields, corrected and supplemented by posteditors.

In summary, it may be said that Soviet researchers have produced good microglossaries, and are likely to produce many more in the near future. In my opinion, the decision to base these glossaries on text is well-founded. It is doubtful, however, that the material gathered with considerable effort for this purpose can be useful for further empirical studies of a syntactic and semantic nature -- that is, for studies in machine abstracting and indexing. In this respect, the larger implications of their work seem to have escaped the researchers.

Linguistic Analysis

MT "systems" in the Soviet Union are comparable with those in other countries. As elsewhere, researchers began with an all-out assault on the problem, attempting to construct analytic rules on an empirical basis, in the belief or hope that the system could be extended, or completed, at some future date. As elsewhere, some of the early demonstrations were performed for purposes of popularization. In my opinion, this phase of MT study is coming to an end, to be succeeded by a phase in which researchers will attempt to learn more about the languages in question. Reference to specific MT systems will perhaps clarify this point.

The most completely described MT systems are those of Belskaya (English-Russian), Kulagina-Mel'chuk (French-Russian), and Mel'chuk (Hungarian-Russian). Belskaya's program may be taken as an example; it is the largest, best reported, and longest in the making. Readers of the 1958 Conference papers are familiar with the threefold process: glossary operation (including lookup, resolution of homography and the syntactic function of symbolic occurrences, identification of part-of-speech of unmatched words, and a certain amount of lexical choice), analysis of the source-language sentence, and synthesis. The approach, and many of the routines themselves, are conventional; the system is essentially an extension and refinement of the earlier Panov-Mukhin version.

I shall not discuss this program in any detail, except to remark that the system does what any operating system can do: performs glossary lookup, identifies idioms (albeit in a roundabout way), performs grammatical analysis of the words in an English sentence, and provides inflection for the Russian dictionary items. Word-order changes can also be made in the Russian sentence, for localized

constructions. It is even likely that the program described by Belskaya can, in theory, adequately translate more sentences than any of the other MT systems which perform these same operations. Does this imply that this program is superior to comparable programs? In my opinion, the answer is clearly negative, for the following reasons. In the first place, as Uspenskij remarked [3], the system is remarkably effective in translating those sentences upon which the algorithm is based; its effectiveness in translating new text remains to be proved. Belskaya reports tests of the system against 3, 000 sentences of new text, using humans who followed the rules, but who did not know English. Such tests are, unfortunately, unconvincing; despite the high quality of sample translations, we are given no indication of the kind and extent of difficulty engendered by new text, nor information about the improvement of the program in light of experience. Has the already extensive system of flow-charts been revised in order to account for exceptions, new conditions, the emergence of new syntactic or semantic categories, etc.? Can the demands on computer programming be met? It may be of more than passing significance that the 1958 version of the system had not been programmed for a computer.

Until we have more information on these questions, we can only surmise that the system is beset with the same problems that beset other empirically-based systems; the rules are not founded on generalizations, or the generalizations are incorrect. For example, Belskaya's routines for resolution of polysemia are still of the type: "Is the first following noun or preposition a member of a certain group, or is the preceding word an adjective?" Ad hoc routines of this type are generally inadequate unless based on an examination of an enormous body of text. If the researcher wishes to learn as he proceeds, his system must be easily adaptable to continuous change. It appears doubtful that any of the three Soviet systems possess this quality; certainly they were not designed with this purpose in mind.

Another characteristic weakness of these systems (at least from my point of view) is the failure to utilize structural information derived from the analytic routines. Syntactic analysis is performed as a part of the translation process, rather than as a step in the research process. By way of contrast, I should like to refer to the approach developed by David Hays at The RAND Corporation.

Beginning with the premise that our knowledge of syntax is inadequate to the task, we constantly seek to extend this knowledge--always, however, through the medium of structure. For a given occurrence, we seldom ask questions like: "What is the following noun?" Rather, we ask such questions as "What is the governor of an occurrence?" or "What is the list of words bearing a certain dependency relation to the occurrence?" Although the approach in both systems is empirical, the methodology is substantially different. In the Soviet systems, new information is hard to come by and arrives in bits and pieces, unrelated and difficult to correlate: it lacks order, pattern, structure.

There are signs that Soviet researchers have become aware of these characteristics in their first MT systems. After all, this work belongs to the period ending in 1957; to the best of my knowledge, it has received much less attention in the recent literature. Whether or not the development of these algorithms has been suspended, it is clear that an appraisal of the current Soviet effort in terms of these systems would be a mistake. The new emphasis is on basic research on the acquisition of knowledge. In the past two years (i. e., in what I have called the second phase of the MT effort) the most interesting and promising work has been done in other areas. For purposes of the present discussion, this area may be called the building of an intermediary language.

An Intermediary Language

Serious thinking about the problem of an intermediary language apparently began in late 1957, as a reaction to the quite dismaying prospect presented by the recently completed algorithms. If the experience of building the English-Russian system were to be repeated for a large number of language pairs, it was obvious that a considerable expenditure of time and money was involved; on the other hand, the success of any of these systems was open to doubt. As readers of the literature know, two approaches to an intermediate language were proposed: by Mel'chuk and by Andreev. The literature with which I am familiar reflects the thinking of these men at an early stage; the differences between their ideas may or may not be relevant today. The differences are summarized by Uspenskij [3] and Mel'chuk [11]. So far as MT research is concerned, the result has been the stimulation of study in language structures and in semantics.

A prerequisite for the building of an intermediary language (IL) is a series of inventories of various kinds for the languages involved. Such inventories would be used whether the IL is thought of as a cross-section or a sum of the various categories. To my mind, some of the most interesting work along these lines has been the isolation and description of syntagmas, or configurations, in Russian and in English. Workers at the Electromodeling Laboratory have been responsible for much of this work: Z. M. Volotskaya, E. V. Paducheva, T. N. Shelimova, A. L. Shumilina, M. M. Langleben [10]; similar studies have been made by T. N. Moloshnaya, Institute of Mathematics [12], and T. N. Nikolaeva, Institute of Precise Mechanics [9] • It appears that these studies in structure have been made without machine aid; nonetheless, a description and comparison of these structure sets will be of great potential value in MT.

Semantics studies range from a mechanized word-count proposed by V. M. Grigoryan [10], to studies of context determinants for resolution of polysemia in German by S. S. Belokrinitskaya [9], and the more theoretical proposals of V. K. Finn and D. G. Lakhuti [10]. Of immediate use in MT are the 20 equivalence determinant classes isolated by Belokrinitskaya for translation of German prepositions; empiricism is here supplemented by generalization. Belskaya's classification of English verbs is interesting for the same reason [9]. There is, in fact, reason to believe that the Soviet researchers have proceeded further in this direction than have researchers elsewhere.

In summary, there is ample reason to believe that Soviet MT workers are now applying themselves to fruitful lines of investigation. The implications of this research to theoretical and applied linguistics are considerable. A significant breakthrough in automatic language translation can be expected if, and when, computer facilities are made available to the large number of workers in the field.

REFERENCES

[1] Panov, D., Concerning the Problem of Machine Translation of Languages, Academy of Sciences, U.S. S.R., 1956.

[2] Kulagina, O.S., "Mashinnyj Perevod s Frantsuzskogo Yazyka . ("Machine Translation from the French") Izvestiya vysshikh uchebnykh zavedenij, Matematika, No. 5, 1958.

[3] Mashinnyj Perevod i Prikladnaya Lingvistika, (Machine Translation and Applied Linguistics) Bulletin of the Association for Machine Translation, No. 1(8), First Moscow State Pedagogical Institute of Foreign Languages, Moscow, 1959.

[4] Tezisy Soveshchaniya Po Matematicheskoj Lingvistike, (Theses of the meeting on mathematical linguistics) Ministry of Higher Education of the U. S. S. R., Leningrad, 1959.

[5] Voprosy Yazykoznaniya, No. 3, 1957.

[6] Voprosy Yazykoznaniya, No. 5, 1958.

[7] Voprosy Yazykoznaniya, No. 1, 1958.

[8] Panov, D. Yu. , Lyapunov, A.A., Mukhin, I.S. , Avtomatizatsiya perevoda s odnogo yazyka na drugoj, (Automatization of Translation from One Language to Another) Academy of Sciences, U. S. S. R., Moscow, 1956.

[9] Sbornik Statej Po Mashinnomu Perevodu, (Collection of articles on machine translation) Institute of Precise Mechanics, AN U.S.S.R., Moscow, 1958.

[10] Abstracts of the Conference on Machine Translation, May 15-21, 1958, U.S. Joint Publications Research Service (JPRS/DC-241).

[11] Vestnik Akademii Nauk SSSR, No. 2, 1959.

[12] Voprosy Yazykoznaniya, No. 4, 1957.