COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

MACHINE TRANSLATION SYSTEMS AND COMPUTER DICTIONARIES IN THE INFORMATION SERVICE. WAYS OF THEIR DEVELOPMENT AND OPERATION

Ivan I. Oubine, Boris D. Tikhomirov

The USSR Center for Translation of Scientific and Technical Literature and Documentation
Moscow, USSR

The paper outlines fields of application of machine translation systems and computer dictionaries in technical translation as well as possible ways of development and operation of computer dictionaries. Interaction between linguists and commercial machine translation system is described

ON APPLICATION OF MACHINE TRANSLATION SYSTEMS AND COMPUTER DICTIONARIES

Translation of scientific and technical literature is assuming these days growing importance in the information service in various branches of science and technology. The volume of translations done at the USSR Center for Translation has grown from 480 thousand pages in 1974 up to 1440 thousand pages in 1980. The same trend is evident in other translation bureaus of the country. Traditional means of translation when the main tools of the translator trade are a type-writer and paper dictionaries do not permit anymore to cope with the avalanche information flow requiring translation. Machine translation systems (MT) and computer dictionaries (CD) are bound to become principal aids to technical translation.

Solving one and the same problem MT systems and CD at the same time essentially differ in structure, aims and technological methods of their application in the information service. MT systems are intended for full translation of texts from one language into another and in principle are aimed at replacing a human translator with a computer working either in automatic mode or in the interactive mode with the human editor. Present day commercial MT systems possess a number of distinctive features. Most important of them are the following: MT systems are bilingual and assymetric (one-way translation), cover one subject field and have high speed of translation. It means that, as a rule, commercial MT systems are designed for translation of texts of a definite subject-field, from the source into the target language and they cannot be reversed or substituted with other languages without a great deal of additional linguistic and programming work. Considerably surpassing the human translator in speed MT systems are substantially inferior to him in quality of translation, that is why machine translation in most cases requires post-editing. Taking into account these features of MT systems as well as the high cost of their development it is expedient to aim MT systems at translation of large volumes of one-subject-field standardized texts. But in practice a great part of technical translations do not meet these requirements. For example, the USSR Center for Translation does translations from 30 languages in 24 subject-fields, and the volume of translations in different languages and subject-fields varies from a dozen to dozens of thousands pages. Thus, languages and

subject-fields with small volumes of translations, texts of advanced complexity or of special importance drop out of the sphere of MT systems employment. To develop MT systems for this purpose would be unprofitable. Nor will it be wise to use MT systems for translation of various kinds of word lists, i.e. catalogues, invoices, lists of spare parts, etc. Such texts should be translated by CD or human translators with the help of CD.

Contrary to MT systems computer dictionaries are designed for supplying a user with the target language equivalents and not for full translation of texts, and are essentially aimed at a dialogue with a human translator to improve his output rather than replace him in the translation process. A technical translator spends half of his working time on dictionary look-up. If we relieve him of this time-consuming chore his productivity will increase essentially.

POSSIBLE WAYS OF CD DEVELOPMENT AND OPERATION

At present a number of computer dictionaries and terminological data banks are in operation in different countries. Most important of them are: the computer dictionary of the Federal Language Service of FRG, terminological data bank EURODICAUTOM of the European Community and TEAM terminology data bank of the Siemens AG. Computer dictionaries vary in structure, computer base, volume of information in the dictionary entry, purpose, subject-fields and other features but at the same time they possess a number of common features which we consider fundamental for CD. Computer dictionaries are mostly multilingual, symmetric, dynamic and cover more than one subject-field.

Taking into consideration wide range and specific character of its tasks the USSR Center for Translation is developing three MT systems (from English, German and French into Russian) and in cooperation with the Moscow State University - computer dictionary MULTILEX. It is designed as a multilingual (Russian, English, German and French) and symmetric dictionary (i.e. query and response are allowed in any of the included languages). Its software is oriented for the use within the framework of the ES computers' system in the following modes:

- single or batch query and response via video display terminal;
- single or batch query via video display terminal, from magnetic tape or cards , and response in the form of a hard copy.

The linguistic form of a query is a very important feature of CD. The possibility of making a query in any text form expands greatly the circle of CD potential users and the range of its applications. CD MULTILEX is designed for users with different language knowledge and can be used for processing standardized texts with certain restrictions in lexicon and syntax as well as texts without any restrictions. In order to meet these requirements CD MULTILEX provides for making queries in any text form for every language included .

There are two basic approaches to solving the text-form-query problem: either to form a search file of word forms or a file of stems or morphs with morphological analysis. In order to achieve optimum performance of a CD the dilemma between a word-form dictionary and a dictionary of stems should be solved with due regard to the grammatical structure of the source language and the type of texts to be processed by CD. The stem dictionary approach is more suitable for synthetic languages with high degree of inflexion (in our case, it is Russian). The word-form dictionaries would better suit analytical languages (in our case, it is English). French and German occupy, in this respect, an intermediate stage between Russian and

English.

In MULTILEX both approaches are used. For English, the word-form-file approach is used. It means that besides the dictionary forms of english entries (which may be uniterms, multiterms or collocations) all their word-forms, spelling variants and standard abbreviations with reference to dictionary forms are stored in the word-form file. Query defining is carried out by mapping the word form realised in the text to the word-forms in the file. For Russian, the stem-file approach is used. Query defining in this case is done by segmentation of the requested word-form into its stem and inflexions. For French and German, we use a combined approach. The response in all the languages is given in the traditional dictionary form.

CD potential is determined to a great extent by the type of information, its volume and composition of the CD entry.MULTILEX entry contains the following information: head word in the dictionary form; part-of-speech symbol; word-forms including spelling variants, abbreviations, or stem/stems; target language equivalents (TE); reliability index of TE; notations which help a user to choose the necessary TE of several registered in the entry; terminological notations which specify the subject-field of the TE; definition of the head word; illustrative phrases; multiterms and collocations which include the head word. As MULTILEX will serve as a new terms bank, a special zone is forseen in its entry for the name of the author of a new term, source and date of registration in the bank. The following information is obligatory for all the entries: head-word, part-of-speech symbol, word forms or stems, subject-field notation, target language equivalents and reliability index.

As a rule, in a new field of science or technology a considerable time is required to form a terminology system and recognize the necessity to issue bilingual terminological dictionaries for this new subject-field. Meanwhile, literature on the subject is beeing translated from one language into another. This fact brings a touch of spontaneity to the process of forming target language terminology system. It leads to the growth of synonymy and polysemy which hampers communication between those working in the new field. The CD makers, as a rule, do not take the responsibility to standardize terminology. Notwithstanding,this problem is not to be overlooked. Should the user trust CD, its every entry should be supplied with reliability index. Considering rigid time limits set for MULTILEX makers they should be provided with a rather simple procedure of determining the reliability index for any TE. Five degrees of reliability have been chosen to suit the purpose. The highest reliability index is given to the TE which comply with the home and international standards. Next in line are the TE borrowed from reliable dictionaries. The TE reliability index depends on the reliability of the source. The lowest reliability index is given to TE which are not registered in any lexicographical source or any other written document. Reliability of a TE source is determined separately for each case. One and the same lexicographical source may have different degrees of reliability for different lexical strata borrowed from this source. The TE reliability index is not a fixed value and is subject to changes through MULTILEX operation.

The users having different tasks and levels of language knowledge, different characteristics of the entry's structure are required. Thus, in traditional lixicography a dictionary entry contains, besides TE, various information about the head word. As a rule, this information is redundant for the professional technical translator who commands profound knowledge of the source language. On the other hand, the entry of a technical dictionary is not as rich and

in most cases, contains practically no additional data besides TE. It is not sufficient for technical specialists whose knowledge of foreign languages is, as a rule, mediocre or poor. This dilemma between the level of foreign language competence of the user and his varied tasks, on the one hand, and the volume and type of data, on the other, cannot be solved in traditional lexicography because of the static nature of paper dictionaries.

CD MULTILEX combines the qualities of general vocabulary and technical dictionaries and solves the above mentioned dilemma by providing the user, in its first stage of development, with the following four variants of response:1) complete entry with the data registered in all zones; 2) entry without collocation zone; 3) entry without collocations and illustrative phrases (i.e. only TE and their notations) 4) one TE only. For entries with more than one TE,the first one is given. We intend to enlarge potential versions of response volume with response variants by subject-field notations, TE reliability index, definitions, etc.

Selection of vocabulary units to a dictionary, CD included, largely depends on its purposes and potential users. CD MULTILEX is designed for off-staff and staff translators and editors, as well as technical specialists. Research shows that the main difficulties for professional translators and editors occur when they are confronted with new, not yet registered in traditional dictionaries uniterms and multiterms, market names of new products, acronyms and abbreviations. Technical specialists consult dictionaries for general vocabulary and general scientific terms as well. Of particular difficulty for them is translation of verbs and verbal collocations.

The first version of CD MULTILEX (English-Russian and Russian-English CD on computer and data processing) is now in operation at the USSR Center for Translation with about 35,000 entries for each language. Multiterms and collocations constitute about 80% of MULTILEX body. The response speed via video display terminal is less than 1 sec. for a single query, for batch processing with alphabetized queries in the form of a hard copy - about 2 min. for a 100 question batch. MULTILEX has proved to be an efficient translation aid. But, in our view, the speeding up of technical translation process should be sought in optimum combination of MT systems and computer dictionaries.

INTERACTION BETWEEN LINGUISTS AND MT SYSTEM IN THE PROCESS OF COMMERCIAL OPERATION

MT systems designed for commercial use require, as a rule, constant upgrading of their linguistic components. This is due to the emergence of new terms, expansion of subject-fields and error-statistic analysis. Consequently production team linguists should refine and perfect the system not only during its development, but in the process of operation as well.

As is known, the development of commercial machine translation (CMT) systems differs considerably from that of experimental machine translation (EMT) systems. The difference is conditioned by the following considerations:

- in EMT restrictions may be imposed on the phrase structure (e.g. syntactic similarity of source and target phrases may be demanded), whereas CMT allows no such restrictions;
- CMT systems handle linguistic data (dictionaries, tables, etc.) considerably larger in volume than that of EMT systems;
- CMT software package must ensure maximum language conversion rate because translation cost depends to a great extent on the time consumed;

- in CMT systems amendment and correction of routines and linguistic components is carried out in the process of commercial operation;
- target texts produced by CMT systems have to be post-edited to achieve the quality required by the customer;
- the form of a hard-copy printout in CMT (page structure, titles, tables, etc.) must not differ considerably from the traditional one;
- EMT systems usually handle a thoroughly pre-edited text, while in CMT a good deal of misprints occur while feeding texts into the computer;

It is evident that CMT systems require a more extensive and thoroughly designed software with due consideration for economic and technological aspects of the problem. A unified software package has been developed in the USSR Center for Translation for the MT systems - AMPAR (English into Russian) and NERPA (German into Russian) - designed for commercial use. Considerable attention has been given to providing production team linguists with convenient means for developing and upgrading the system.

Since inter- and post-editing are used in the systems to improve the quality of translation, the process is divided into several stages: initial processing, inter-editing, language conversion, post-editing, and target text printout. Stages in their turn, are split into steps comprising the most essential algorithms of the system. Steps that are expected to undergo frequent alterations, are composed of step subroutines, each realizing some specific parsing or synthesis procedure. Step subroutines consist of operators performing definite linguistic functions. A special programming language for linguistic procedures has been devised in order to give linguists a means of maximum direct participation in development and improvement of the MT systems.

Information files of the MT systems include dictionaries, tables, source and target texts, and a corpus of information cells where essential information about words or groups of words is stored. Subroutines compiled by production team linguists are fed into the computer and converted into a form convenient for futher use. In the process of translation they are called up from disk storage into the MOS memory and executed by the interpretation program. The assembly language has been used by programming engineers to develop programs not subject to alterations (text feed-in, dictionary look-up, monitoring, etc.) with the aim of minimizing operation time.

The modular structure of the software package, where an algorithm is divided into several procedures each beeing executed by a separate program module, considerably facilitates program compiling. Information files (subject dictionaries, tables) are also characterized by modular structure. Due to relative independence of modules, the software package is flexible enough and can be easily revised by introducing new modules, correcting or deleting the existing ones or reordering the sequence of their operation; that is, the software makes it possible to generate different versions of the MT system during both running-in and commercial operation.

The basic operation modes of the system are:file maintenance, monitoring and work. File maintenance mode serves to create, amend, correct, and print files used in processing source language texts. File maintenance software package developed for this purpose acts autonomously of the text processing programs. The work mode implies mass translation of texts by the operational version of the system in which the sequence of steps is fixed and work data files are used. The choice of subject-field dictionaries is determined by the control information accompanying each text. In addition to a target language text, some intermediate information is printed out in the work mode

concerning errors, missing words, ıtradictory situations, etc.

In some cases the intermediate inf'or.,tion is sufficient to determine the nature of the errors. In other cas·s,linguists can obtain additional information by rerunnin{; the defective fragments in the monitoring mode. While the same operational version of the system as for the work mode is used, a highly detailed selective printout is obtained by typing in special instructions. At request the following data can be printed out: the input text with a reference grid helping to select required fragments; the corpus of information cells accompanied by the corresponding source language and target language words or collocations (the latter after translation); information on subroutine operators and their sequence of work; alterations produced by subroutine operators into the information cells; entries selected from the source dictionary to match text word-forms; target dictionary entries selected for synthesis; source and target language sentence pairs. The nature and volume of printed information is determined by the linguist depending on the problem facing him.

The MT systems have two kinds of files: operational and upgrading. At the initial stage upgrading files are identical to the operation ones. Then, on the basis of error analysis, linguists correct and amend the upgrading files. As a sufficient number of alterations is introduced into the upgrading files, a duplicate version of the system is generated and all amendments are checked and tested in the monitoring mode. After a thorough checking the upgraded files replace the old ones in the operational version of the system. The information in the operational files is rearranged so as to avoid degradation of the processing rate.

As a result, production team linguists have the following opportunities: to take part directly in developing and perfecting program modules; to obtain information on missing dictionary entries and typical errors; to promptly determine the nature of error and locate it; to produce diverse versions of the system by introducing new and altering the existing modules and their sequence without interfering with commercial operation of the system; to check the applicability of newly produced version for singling out the most efficient one to be input as the operational version; to monitor the state of linguistic data and program modules and correct them if necessary; to intervene in the process of translation at different stages in order to introduce corrections into the text being processed.

Linguists interact with operational and service programs and define operation modes with the help of specific control language instructions. Instructions are typed in before or during the performance.

Thus, the described software package provides production team linguists with the means to develop and adjust step routines, to compile and upgrade linguistic data files, and to monitor the MT system without programmers' assistance.