Some Specific Features of Software and Technology in the AMPAR and NERPA Systems of Machine Translation

B. D. TIKHOMIROV

All-Union Centre
for Translation of Scientific
and Technical Literature
and Documentation.
ul. Krzhizhanovskogo, 14,
117218 Moscow. USSR

The author considers specific features of software and methods of operation applied in two machine translation systems based on a common linguistic approach and intended to translate texts from English and German into Russian on an industrial scale. He describes the software structural pattern and linguistic tools used in the systems, as well as structures of specialised information files and their interaction in the course of translation. He discusses forms of utilisation for machine translation systems operating on an industrial scale.

INTRODUCTION

In creating the AMPAR and NERPA machine translation systems (from English into Russian and from German into Russian, respectively) at the All-Union Centre for Translation of Scientific and Technical Literature and Documentation, unified software oriented to commercial operation was developed [1, 2]. Much effort was made to facilitate the work of system linguists in the course of the development and operation of these systems (3).

Main specific features of software and technology are as follows:

— break-down of the translation process into a number of stages,

— use of a specialised programming language for specific algorithms of machine translation,

— two-stage organisation of information files,

— availability of topical information files for various subject fields,

— use of a specialised process-control language to specify information-file input/output instructions or information-file processing modes,

— possibility of obtaining system-operation results from any stage in a form convenient for their analysis,

— possibility of reorganising the structure of the system (creation of different versions to select the most efficient one),

— possibility of system generation with the prespecified set of functions and topical files.

SOFTWARE STRUCTURE AND LANGUAGES

Since the AMPAR and NERPA systems use inter-editing of intermediate results and post-editing of target texts to increase the translation quality, the process is divided into a number of stages: initial processing, inter-editing, automatic translation, post-editing and target text printout. In turn, a stage may be split into a number of steps, each implementing some essential algorithm of the system.

Steps liable to frequent alterations consist of the step subroutines (schemes), each executing some specific analysis and synthesis algorithm. The step scheme consists of statements, each executing a certain linguistic operation.

To achieve the maximum participation of the linguists in the process of creating and modifying the specific algorithms of the system, a special programming language oriented to a certain data structure has been developed. Information files contain dictionaries, tables, source and target texts, and a file of information locations to store and modify information on & word or word combination required in the course of translation. Throughout translation the schemes are called from the external memory into the main memory and the specific algorithms are implemented using an interpreter program.

Subroutines which are not modified during the system operation (the source text input routines, dictionary look-up routines, monitor, etc.) are developed by the system programmer using the assembler language of the ES Computer System to minimize their running time. Use of a modular principle for creation of the software structure in which an algorithm is divided into sufficiently small algorithms and each specific algorithm is program-implemented as a separate module ensures a significant simplification of the programming process. The modular principle is also applied to the information files (subject-field dictionaries, tables).

Due to a relative independence of modules, the software system obtains enough flexibility to allow easy updating of programs and information files by inclusion of new modules, by updating (or deleting) modules or changing their sequence, i. e. it is relatively simple to generate various system versions by debugging and updating it when it is in operation.

The main modes of the system operation are: file maintenance, monitoring, and work modes,

The first mode allows you to create, add, update, and

list files that are used to process the source text in the monitoring and work modes. To maintain files, a utility package independent of work routines has been created. The work mode is used for mass processing of texts by a work version in which the sequence of all system element operations is fixed and the work data files are employed. Choice of subject field dictionaries is determined by the control information accompanying each text. In addition to the target text, intermediate information is provided in the work mode on certain errors which occur during text processing, words not found in dictionaries of the system, contradictory situations, etc.

In some cases the intermediate information is sufficient to determine the nature of faults. In other cases, to obtain additional information, the system linguist can have the faulty text fragments reprocessed in the monitoring mode. The same system version is used as in the work mode, but a selective listing of the system operation at any prespecified section of the text with a high degree of detail (with accuracy of up to operation of an individual module statement) is set up using a special directive.

The machine translation system employs files of two types; operational and upgrading. Originally, the upgrade files are fully identical to the operational files. On the basis of information about the nature and location of an error, the linguists correct the specific algorithms and/or information files which constitute part of the upgrade file, as well as create new versions of the individual schemes. After changes have been introduced into the upgrade file, an upgrade system version is generated in which the correctness of the introduced modifications is checked out in the monitoring mode. Operational files include only those elements of upgrade files for which work quality has been checked out in the monitoring mode.

As a result, the system linguists have an opportunity:

—to participate directly in development and debugging of modules implementing specific algorithms of the system,

—to identify missing dictionary entries and typical errors occurring during text processing,

—promptly localize an error and determine its nature,

—to create versions of the system without disrupting the commercial operation of the system and to maintain its operational version intact; each new version may include new and/or modified program and information modules or may change the order of their operation,

—to verify operation of created versions using textual material of a large volume in order to select the most efficient version and to introduce it into the operating file as the operation version,

—to monitor the state of information files and program modules and to ensure their rapid updating.

The specialised process control language (PCL) implemented as a set of directives according to which the program modules of the system perform certain operations is used as a language for interaction with the system and organisation of the intermodular links inside the system. The set of directives, called an instruction, is entered into the computer before solving a task or in the process of its execution. In addition, each program module may issue an instruction that is to be processed by the called module. PCL allows us to alter the standard order of the programs, to specify different printout modes, and to correct the information files.

VERSIONS OF MACHINE TRANSLATION SYSTEMS

The AMPAR and NERPA systems have been developed as multifunctional machine translation systems (MMTS) intended for use in large translation organisations [4] and must provide:

— translation of polythematic documents,

— adjustment to any form of input information including information retrieved from data banks or as software descriptions on magnetic tapes,

— input and processing of information having a sophisticated structure,

— inter- and post-editing in the interactive mode,

— conducting broad investigations in the field of lexicography,

— prompt correction and upgrading of information files,

— generation of MMTS with prespecified sets of functions and subject field files.

It is obvious that MMTS must be maintained by the system linguists and programmers who are fully aware of the particularities of the multifunctional machine translation systems. To hand the multifunctional machine translation system over to any other translating organisations, it is expedient to generate a simplified version of the multifunctional system, a specialised MMTS.

Any specialised MMTS must have fewer functions than any general-purpose MMTS; its information files must be oriented to specific subject fields. Provisions should also be made so that the system may be maintained by personnel of not very high qualification on any computer configuration.

TWO-STAGE ORGANISATION OF INFORMATION FILES

Such factors as speed of translation and ease of the linguists' interaction with the system during its development and operation are of great importance for the commercial MMTS.

Obviously it is impossible to select the information representation form that might be equally suitable for a human being and a computer. As a means of settling this discrepancy, provision is made to ensure a two-stage organisation of the information files in the AMPAR and NERPA systems. This implies that two files, a linguistic information file (LIF) and a machine information file (MIF), as well as converters of information from one form to another are created. Both files are stored in the computer memory, the LIF being amended and updated using a mini-computer. Each file is divided into a number of subfiles by functional and technological characteristics.

Information is stored in LIF in the form of words and word combinations in the source and target languages with the aid of convenient mnemonic codes and

a programming language of specific algorithms. L1F consists of subfiles of separate words and word combinations, ambiguous words, and grammar.

In MIF information coding technique and information arrangement are selected with a due regard for the most efficient data processing by the system programs. MIF contains subfiles of the source and target dictionaries, homography tables, word combinations, translation equivalents, subprograms for translation of any compound word combination and ambiguous words, as well as grammatical subprograms.

Each element in LIF (a dictionary entry or a scheme) is accompanied by keys indicating that the entry belongs to a specific subject field. The said keys also indicate an entry creation or update date, etc. A service routine periodically selects an entry matching the prespecified key from LIF and forms an update file to be further handled by the converter, mapping each entry of LIF onto one or several entries of MIF and writing it into the MIF subfile.

Use of the two-stage organisation of information files allows us:

— to simplify considerably the linguists' work with the system due to elimination of information having a non-linguistic nature and also due to the fact that the linguists are relieved from taking into account all links that arise when information on each word is embedded in several subfiles of the system,

— to decrease the number of errors in MIF which are associated with coding mistakes and data transfer onto the machine-readable medium,

— to form several subject-field MIFs using a single LIF,

— to exchange information with automated dictionaries.

SUBJECT-FIELD INFORMATION FILES

Each element of a linguistic information file in MMTS may consist of several fields related to different subject fields and containing various information. Each such field has a subject field code.

While MIF is formed, entries that have identical subject codes are grouped into separate subject-field blocks (dictionaries, tables) in each MIF subfile. The aggregate of blocks related to one subject field is called the subject-field information file (SFF). A special role is played by the base information file (B1F), or the common vocabulary file.

When translating a text related to a certain subject field, BIF and one or several SFFs are used. Order of their operation is specified when post-editing. As a rule, BIF has the lowest priority. At each stage, the specified subject-field blocks associated with the appropriate MIF subtile are selected, and information from these blocks is used in accordance with the prespecified order.

For instance, at some stage the words from the source text are first matched against the subject-field dictionary having the highest priority. Then the words missing in the subject-field dictionary are matched against the base dictionary (common vocabulary dictionary).

Main advantage of the modular generation and use of information as compared to utilisation of a single file, from our point of view, lies in a complete independence of all subject-field files. This allows several independent groups of linguists who work in different subject fields to participate in files extension. Besides, making the subject field more narrow simplifies creation of files due to decrease of lexical-grammatical homography and lexical ambiguity, which enables us to enlist specialists unfamiliar with the particularities of the system linguistic support for creation of files.

Another important consideration is the fact that despite the enlargement of the entire information file (due to duplication of information in various SFFs), when a translation is made, only a part of the common file is read into the main memory of a computer, thereby decreasing the translation time.

CONCLUSION

The program package which was originally developed for machine translation from English into Russian, turned out to be basically applicable to machine translation from German into Russian, as well, thanks to the general-purpose features used in it. The specific features of German required insignificant modification of the program complex, development of some new program modules, alteration of the source dictionary organisation, and introduction of new stages.

Operation of the AMPAR and NERPA machine translation systems confirmed the correctness of the adopted software and technological decisions.

REFERENCES

1. Marčuk, Ju. N; Tihomirov, B. D.; Ščerbinin, V. I. Ein System zur maschinellen Übersetzung aus dem Englischen ins Russische. In: *Sonderdruck aus Automatischer Sprachübersetzung.* Darmstadt, 1982, 319-336.
2. Marčuk, Yu. N.; Vlasov, A. N. Some principles of computerised translation from German into Russian. *Fremdsprachen,* 1980, No. 2, 91—99 (in Russian).
3. Oubine, I. I.; Tikhomirov B. D. Machine translation systems and computer dictionaries in the information service. Ways of their development and operation, *Proceedings* of *the Ninth International Conference on Computational Linguistics.* Prague, 1982, 289—294.
4. Tikhomirov, B. D. Possible ways and forms of machine translation application at an industrial scale. In: *International Seminar on Machine Translation. Paper summaries.* Moscow. VTsP Publishers, 1983, 217—218 (in Russian).