Machine-Aiding of Translation Processes by User-Controlled Dictionary Look-up

GUNTER NEUBERT

Reviews automatic dictionaries of various types. The usefulness of terminology data banks in different spheres of human activity is illustrated. The operation and various applications of the Dresden electronic specialized dictionary are described.

It is very well known that machine translation being the long-term offer of linguistic engineers for the full-scale automation of the translation process cannot easily and in a short time be implemented in all fields and in all languages. One of the potential by-passes, simultaneously constituting a necessary step towards machine translation, consists in automating the operation of dictionary look-up.

1. Comprehending the special terms of the text to be translated and selecting the adequate term from sets of equivalents listed in the dictionary entry, or constructing a likely equivalent by processing the different terminological informations collected, present most of the problems of the translation process and demand the greatest proportion of expended time, sometimes up to 80 per cent of it. For that reason, the preparation of a data file containing terminological information and utilisable with programs is obviously most promising in regard to machine-aided translation. In addition to the consequent concentration of lexicographical power, such a terminological data file, or data bank as it is more frequently called, allows various other data handling operations to be automatically performed, in this way overcoming some of the shortages of traditional dictionaries.

For example, it is feasible to derive special lists of terms, or vocabularies, containing all the stored terms with specific features, such as being used in a special field or in the context of a special product; and/or being a member of a special category of concepts like instruments, tools, machines, procedures, etc.; and/or pertaining to a given character string like English pump, German Pumpe or Russian насос. Interpreters may use such vocabularies for preparing themselves for a congress or discussion, translators may complete their terminological work...
ties for international data interchange and a higher degree of accessibility for all kinds of users will be available.

The programs perform data input and updating operations, both of general and of specialised types, data retrieval, selecting operations, and editing procedures. Besides, linguists are given, facilities to analyse German word compounds for their morphmatic and lexical constituents, and to morphematically decompose Russian words.

1. These latter routines are considered to be most important for enhancing the data bank output in the case of machine-aided translation. It is well known that a considerable amount of the questions put to a dictionary will not be answered when based on a character-by-character comparison of the Searched Term (ST) with the set of Compiled Terms (CT). As some of our investigations clearly show, the amount of directly answered terminological questions by even a high-level special dictionary does not exceed some 30 per cent. In addition to this fact, which will hardly be overcome by strengthening terminological efforts, as it results from the fast and progressively growing development of science and technology, translators sitting in front of a common displaying screen cannot survey the alphabetical or semantic environment of their ST. They are dependent upon their experience in using dictionaries and, besides, have to type question after question on the terminal keyboard, until they finally find sufficient auxiliary information to form a likely term in the target language.

Various techniques of matching ST and CT are known, and are being provided by producing terminological data banks. As the most natural way, searching difficulties can be reduced by eliminating such character differences as resulting from capitalisation, diacritic signs or the German 'ß', and word differences due to the somewhat arbitrary hypenisation of the English language. These difficulties can be reduced further by cutting off one, two or more of the last characters of the ST after having compared them with the characters or character strings respectively, given a list of possible endings in the language concerned. In this way, differences between word forms like hydrocarbons as ST vs. hydrocarbon as CT, or reversed, or Verdampfen vs. Verdampfung, or vsasyvani as vsasyvat' are ignored. One of the most widespread techniques is based on the principle of longest-possible match, i.e. to base the answer on the CT that is the longest continuous character string identical with the initial characters of ST.

All these and some other related techniques not mentioned here have the disadvantage of not enabling users to control the generation of the term substituted for an ST not found in the data bank. Users may command the execution of any of the following procedures:

(a) matching the first \( n \) characters of ST;
(b) matching the last \( n \) characters of ST;
(c) matching the first \( n \) characters of an index word of the ST in the case of multiple word compounds; or
(d) matching the last \( n \) characters of that index word.

In general then procedure (a) provides users with the equivalents of the determining constituents of ST, while procedure (b) gives the equivalents of the determined constituents, at least in the case of word compounds with the corresponding word order. The reader can easily imagine modifications for other cases of word order.

In order to avoid a time-consuming count of ST characters, users may also prescribe the value of \( n \) as the number of residual characters of ST to be excluded from the comparison, by adding a negative sign preceding \( n \).

Without having repeated his ST the user receives a list of CT corresponding to the enlarged retrieval conditions printed out on the display, and decides which of the entries is or are to be put out completely. Thus, he can collect terminological information step after step, if necessary, to form the desired term.

The above mentioned procedures enable users to specify searching policies dependent on the actual ST. However, it is evident that in many cases the programs by itself, or more exactly the linguistic algorithm being the base of the program, can generate auxiliary terms that may be considered to be probable synonyms or constituents of the original ST. In all languages concerned, groups of synonyms or term variants exist like the German Wiederholungsstart and Wiederholstart, the English seat valve and seating valve, or the Russian sglazivajuščij kondensator and kondensator dlja sglazivanja, which can be mutually transformed following language-bound rules which, however, have not been sufficiently defined up to now. Just as well, programs can be made to dismember word compounds, which are especially relevant in German, into their constituents. For those transforming or dismembering routines, morphematically analysing algorithms are the necessary fundamentals. At present, they are in the stage of draft for the German and Russian languages.

In the case of non-direct man-machine dialogue, users can be given the possibility to pre-call the earlier mentioned match enlarging procedures to be executed during the processing of a list of ST only in the case of non-existence of the original ST in the stores of the data bank. However, use of this principle is justified only in lexicographically well-compiled fields to avoid a considerable 'swelling' of the output lists.

In any case, each step towards solving the problem of an effective retrieval of terminological information will also be a step towards solving the problem of continuously filling the terminological stores, for satisfied users will readily undertake the additional task of contributing their terminological by-products from translating to an easily accessible file.

REFERENCES