Igor Boguslavsky

A Bi-directional Russian-to-English Machine Translation System (ETAP-3)

Abstract.

The bi-directional machine translation system ETAP-3 developed by the Computational Linguistics Laboratory of the Institute for Information Transmission Problems of the Russian Academy of Sciences is a part of a larger system - a multi-purpose linguistic processor. Its main options are:

1. Russian-to-English machine translation.
2. English-to-Russian machine translation.
3. Natural language front end for databases (with the Russian input).
4. Natural language front end for databases (with the English input).
5. Computer-aided teaching and learning of lexica.

Theoretically, the whole project is based on the Meaning ↔ Text theory of language originally suggested by Igor Mel'chuk and later elaborated lexicographically by Jurij Apresjan into the theory of integral description of language. The computer version of the Meaning ↔ Text theory built up by the authors follows rather closely the multilayer structure of its prototype. In fact, it is a linguistic knowledge database providing tools for processing written texts (mostly scientific, technical, or documentary) in unconstrained natural languages for diverse subject domains and diverse processing tasks. The linguistic knowledge is presented independently of the algorithms which make use of it and is not geared to a particular subject domain, to a particular processing task, or to a particular natural language. Due to these properties it may be thought of as a multi-purpose and multi-lingual computer system of language processing which can, in case of need, accept new working languages, new subject domains, and be extended to new problems. Being independent of the algorithms, it is directly observable and can be easily updated,

The ETAP-3 system is intended for translation of scientific texts from the domains of electrical engineering and computer science between English and Russian. It is a basically revised and improved version of the unidirectional English-Russian MT system ETAP-2, developed by the authors in 1982-1985.

The system is based on combinatorial dictionaries of English and Russian, used in both directions of translation, which count over 16,000 entries each, complete grammars of English and Russian (both morphology and syntax), providing for the coverage of all the grammatical forms of the working languages and almost all the syntactic constructions that occur in the processed texts, and vast sets of transfer rules. The dictionaries ensure a satisfactory coverage of texts for experimental operation, though they are not sufficient for commercial purposes. The system handles unprepared natural texts sentence by sentence, taking 6 to 20 seconds per an average sentence 20 to 30 words long. The translation process requires no human intervention, although limited technical pre-editing is done and some post-editing is desirable (in many cases, to choose between two or more alternative translations of an ambiguous sentence which the system may offer at the request of the user).

The four largest components of the system - the morphological and the combinatorial dictionaries of English and Russian - are designed in such a way as to be used in both directions of translation. In English-to-Russian translation, the English dictionaries serve for the morphological analysis and parsing, while the Russian dictionaries arc used for the syntactic and morphological synthesis. In the Russian-to-English option the roles of the

dictionaries are swapped. The other components (parsing, transfer and syntactic synthesis rules) are used in one option of translation only but they arc based on the same linguistic principles and use the same formats of linguistic knowledge representation.

The system is implemented on the Micro VAX 3100 computer. All the software is written in C. The system contains a set of software tools aimed at facilitating the process of dictionary and transfer rules composition.

A textual database is in preparation which will contain parsed texts belonging to a certain subject domain. Joined to the parser, this database will allow to disambiguate ambiguous sentences on the basis of the texts parsed earlier.

The work is being done on extending the system to three more languages - French, German and Korean.

Name, title, function, affiliation:

Boguslavsky Igor, Dr. Sc.(Linguistics).
Head of the Computational Linguistics Laboratory, Institute for Information Transmission Problems, Russian Academy of Sciences.
19Ermolovoj St, GSP-4, Moscow 101447 Russia