THE JaRAP EXPERIMENTAL SYSTEM OF JAPANESE-RUSSIAN AUTOMATIC TRANSLATION

Larisa S. Modina, Zoya M. Shalyapina

Institute of Oriental Studies, Russian Academy of Sciences,
Rozhdestvenka str., 12, 103753 Moscow, Russia

Abstract

The paper is the first report on the experimental MT system developed as part of the Japanese-Russian Automatic translation Project (JaRAP). The system follows the transfer approach to MT. Limited so far to lexico-morphological processing, it is seen as a foundation for more ambitious linguistic research. The system is implemented on IBM PC, MS DOS, in Arity Prolog (analysis and transfer) and Turbo Pascal (synthesis).

1 Theoretical background

The development of the JaRAP experimental system was preceded by a long period of purely theoretic research into various aspects of natural language and its functioning in translation (see, e.g., (Shalyapina,1980a,1980b,1988)). Some of the basic principles which have evolved from this research may be summarized as follows.

(1) The most adequate scheme for simulating human translation activity is doubtless the transfer one.

(2) The level of transfer and the volume of structural and semantic information explicitly represented at this level should be determined experimentally as a compromise between the demands for translation adequacy under the given conditions and the advantages of "shortcuts" permitted by the superficial correspondences between the languages concerned.

(3) Semantics is not in itself a level of linguistic representation, but rather part of linguistic description at any level of representation of linguistic units.

(4) In its semantic aspects, syntax is dependent on lexicon to a greater extent than vice versa.

(5) A model aimed at faithful simulation of linguistic performance should make explicit use of the factor of linguistic normativity, this being, at least in prospect, a building block for "self-tuning" functions as an analogue for human learning capabilities.

An approach best suited for effectuating these principles seems to be that of relying on a lexicon-oriented lingware framework of a special kind. Within this framework, entries of a uniform structure may be provided, besides lexical units, also for morphological categories, function elements (including punctuation), and all kinds of grammatical features, while syntagmatics of all levels may be presented in terms of valencies of those levels, assigned to the corresponding lexical or grammatical units in their entries.

The JaRAP experimental system is meant to incorporate this approach.

In accordance with the transfer scheme of translation, the system is made up of three major components: the Japanese analysis component, the Japanese-Russian transfer component, and the Russian synthesis (generation) component. It is implemented on IBM PC, MS DOS, its programming tools being Arity Prolog for analysis and transfer, and Turbo Pascal for synthesis.

2 The current version of the JaRAP system

At present, the JaRAP system does not go far beyond the initial lexico-morphological level of text processing (though some provision has already been made for further stages of its development - see Sec.3).

The **analysis** component of the system performs so far three main groups of operations: ***segmentation*** of the input Japanese texts into graphico-morphological (GM-) elements (stems and suffixes of Japanese words); ***processing of translationally idiomatic (TI-) combinations*** of GM-elements; and ***lexico-morphological (LM-) analysis*** of the resulting sequence of GM-elements and their TI-combinations.

Segmentation is accomplished in two steps. First, the input text (= the input sequence of kana and kanji kodes) is broken up into fragments by *contextual delimiters* certain to denote word or morph boundaries (e.g., punctuation marks, the occurrence of a katakana symbol after a hiragana one or vice versa, etc.). Then the fragments obtained are segmented into GM-elements by means of dictionary search. The resulting GM-elements are represented by the reference numbers locating their dictionary entries in the database used. For segmentationally ambiguous fragments, all possible segmentations are formed. If dictionary search is unsuccessful, the program draws on an auxiliary index of separate graphic symbols, so that "unknown" words can still be processed (and if they are composed of kanji, be even provided later on with a translation of sorts).

The processing of TI combinations of GM-elements is partly necessitated by the fact that fragment boundaries may sometimes separate the components of a compound word, like

読み | 取り

so that these components have then to be joined together by a special procedure. The same procedure is used to locate multi-word combinations similar to single GM-elements in that they have idiomatic translations and do not allow of variations in their internal structure (this is often he case with terminological expressions). TI-combinations are searched for as sequences of reference numbers identifying the GM-elements they are composed of. When found, they are replaced each by a single reference number - that of the entry for the TI-combination as a whole, and are subsequently treated in the same way as individual GM-elements (with some reservations mentioned in Sec.3).

LM-analysis of a sequence of GM-elements examines, for each of them, all of its alternative lexico-morphological interpretations, or LM-elements contained in its entry, with the aim of integrating the LM-elements corresponding to adjacent GM-elements into acceptable morphologtical (M-) representations of Japanese word-forms. The acceptability of these is established by checking each M-representation, as soon as it is formed, for the co-occurrence restrictions its elements may impose on each other and on the elements of its immediate contextual neighbours. This also serves for disambiguation, as all the LM-elements that cannot be used to form an acceptable M-representation of a word-form in the given sequence of GM-elements, are filtered out.

To optimize processing where alternative paths of analysis are concerned, all analysis procedures are organized so as to limit separate processing of such alternatives only to the subpaths responsible for the differences between them. If some subpath is the same in two or more of the alternative analyses, it is processed just once, and the result is used for all the corresponding alternatives.

The bulk of the **morphological description** used in LM-analysis is of a *valency-based* type (an exception being the morphonological - or, rather, morphographical - alternations: the 10 metarules representing such alternations are incorporated in the segmentation procedure). The morphological valencies are mostly assigned to suffixes, while stems (verbal or adjectival) act as fillers. The co-occurrence restrictions imposed by the elements of a word-form on those of its adjacent word-forms are described in much the same way (the only difference being that in this case the data to be checked is assigned to stems at least as often as to suffixes). This helps to make word-boundaries transparent, if necessary, to morphological valencies, so that the borderline between morphology and syntax loses something of its traditional rigidity.

Transfer operations at the lexico-morphological level are limited at present to those of replacing the elements of the Japanese M-representation obtained from analysis, by their Russian equivalents, and shifting, where necessary, the Russian morphological categories that may appear as a result of such replacement, from the positions they initially occur in to their appropriate word-forms. Sometimes this involves skipping a number of intermediate elements, such as auxiliaries, brackets, etc.

Besides lexico-morphological transfer, we have by now implemented some very simple syntactical analysis-and-transfer operations based on the most general correspondences between Japanese and Russian structural and word-order information. This is only the very first step to the syntactical transfer component we are planning, but the operations implemented are already sufficient to provide adequate Russian translations for Japanese sentences containing no embedded clauses, lexical ambiguities, or other difficult linguistic phenomena.
Thus, the sentence:

日露機械翻訳システムは
多くの人に必要である．

Nichi-ro kikai hon'yaku shisutemu wa
ooku no hito ni hitsuyoo de aroo

is translated as:

Система японско-русского машинного
перевода является, по-видимому,
необходимой многим людям.

The information database used in the analysis and transfer procedures is organized as an indexed list of dictionary entries for individual GM-elements, TI-combinations of GM-elements, and grammatical features (classes) of LM-elements. To speed up dictionary search, the database is provided with an index organized as a superposition of balanced trees.

Each entry (presented in the database by a Prolog term) constitutes a list of entry zones confined each to one type of linguistic information. A separate zone (identified by the corresponding label) is used to specify, e.g., the graphical representation of the GM-element described, its structural (lexico-morphological) representation; the list of its grammatical markers; each type of restrictions imposed on the elements filling its morphological valencies, etc. The overall set of entry zones is the same for all types of entries, though each entry contains only the zones relevant to the element described.

At present, the database includes over two thousand entries.

Special emphasis has been placed upon providing the system with *efficient means of updating linguistic information.* The environment built for this purpose is called VOCOPS ("VOCabulary updating OPtionS").

The VOCOPS environment allows the user to add, delete or replace all types of dictionary entries or zones within them in a highly interactive mode. VOCOPS checks the updating information for its formal accuracy and for its compatibility with the information already contained in the current database. It then proceeds to warn the user of those consequences of his updating operations which otherwise might have been overlooked, and to indicate the inaccuracies or inconsistencies detected. If possible, it also suggests the likely ways of their correction. Among other things, VOCOPS keeps watch on the correspondence between the entries for individual GM- (and LM-) elements and those for their TI-combinations. E.g., if the user wishes to delete a GM-element which forms part of some of the TI-combinations present in the database, VOCOPS lists these with a warning that they will also be deleted.

The Russian synthesis component is constructed as an independent subsystem, complete with a database of its own. Its functions include both morphological generation and some aspects of syntactic processing. Here we will not discuss it an any length, because there is a separate paper devoted entirely to this component (Kanovich,Shalyapina,1994).

3 Development work under way

Implementing the most basic (however simple) of the linguistic functions needed in translation, the current version of the JaRAP system constitutes the necessary foundation for further developments. Both its database and its programming software are structured to accept any new components (new zones of the dictionary entries, new programs, etc) without impairing those already functioning. The VOCOPS updating subsystem is also general enough to be easily tuned up to new types of linguistic data as soon as they are included in the system.

Moreover, even in its present form, the JaRAP system comprises some specific features meant for more advanced linguistic processing.

Thus, among the grammatical markers assigned to the Japanese LM-elements in the current database are a number of those to be used in syntactical analysis.

Entries for TI-combinations of GM-elements include specification of their syntactically and semantically dominant components, for use in processing parallel constructions and anaphora.

The list of the Russian equivalents for an LM-element includes, wherever desirable, different parts of speech, the choice between them to be effected by the syntactical transfer.

The synthesis component is designed to accept syntactically weighted representations of Russian word-forms, etc.

Now that we have built the basic groundwork, labor-consuming as it is, we are taking up these, more ambitious tasks.

As the Japanese-Russian pair of languages is virtually unexplored in its machine-translation perspective, our immediate efforts are being focussed on determining the reasonable minimum of grammatical knowledge of Japanese necessary for obtaining intelligible Russian output for unadapted (un-pre-edited) Japanese input.

References

[1] Kanovich, M.I., Shalyapina, Z.M. (1994) The RUMORS system of Russian synthesis (submitted for COLING 94).

[2] Shalyapina, Z.M. (1980). Automatic translation as a model of the human translation activity. *International Forum on Information and Documentation, v.5,* No.2, p.18–23.

[3] Shalyapina, Z.M. (1980). Problems of formal representation of text structure from the point of view of automatic translation. In *COLING 80. Proceedings of the 8th International Conference on computational Linguistic.* Tokyo, p.174–182.

[4] Shalyapina, Z.M. (1988). Text as an object of automatic translation. In *Tekst i perevod.* Moscow, p.113–129 (in Russian)