The Role of the Lexicon in a Semi-Direct MT System

Karel Oliva

Abstract

This paper gives information about the contents and organization of the dictionary in the Czech-to-Russian machine translation system developed at Charles University, Prague. It is particularly concerned with the nature and also the use of information contained in a dictionary item. The paper also briefly describes the overall architecture of the system, the role played in the system by the dictionary, and points out the advantages of the major redirection of linguistic information to the dictionary in this system.

Introduction

The aim of this paper is:

a. to present briefly the design of the Czech-to-Russian machine translation (MT) project, being developed cooperatively at Charles University, Prague, and the Research Institute for Computing Machinery, Prague

b. to give information about the contents and organization of the dictionary in this project

c. to elucidate the nature of information contained in a dictionary entry in this project

d. to point out the practical advantages of the major redirection of linguistic information from grammar to dictionary, made in the system.

General Design

The underlying idea of the Czech-to-Russian MT system discussed is that of the transfer approach (cf. Slocum 1985 or Hutchins 1986 for more detail); however, the practical requirements (the system is purely production-oriented) led to sweeping changes in the "incarnation" of the classical transfer philosophy. In the absolute majority of cases these changes were made possible by the relative closeness of Czech and Russian, particularly in their syntactic structure: this enabled compressing the transfer phase to a minimum which was then, on implementation, divided and shifted either to the Czech analysis module or to the Russian synthesis module, thus making these modules mutually dependent. (This is actually the reason for calling the approach "semidirect", to distinguish it both from the "pure" transfer approach on the one hand and from primitive "word for word" translation on the other).

Dictionary: its Status, Contents and Organization

Not only in practical systems in the field of computational linguistics, but even in theoretical frameworks of language description it is quite difficult to draw the borderline between *grammar* and *dictionary*, i.e. between systemic and lexically-bound information: in any working system this borderline must (and in any theoretical framework, it should) be drawn clearly, leaving no space for inconsistency, discrepancies and omissions. On the other hand, the decision *where* (as opposed to *how*) to draw this borderline is quite arbitrary — in extreme cases it is possible to keep all information in just one component (and to suppress the other one completely). In the project described it was decided to store all information which is bound to a particular lexical unit with this unit in the dictionary, even in those cases where this information concerns some purely syntactic properties or the behaviour of the unit during the translation process. The grammar, then, contains rules of basically two kinds: first, the most general rules of Czech and Russian syntax (e.g. rules concerning subject-verb agreement), and, second, highly schematized rules processing the concrete pieces of information projected to them from the particular lexical units stored in the dictionary (e.g. rules filling the slots of case frames or rules carrying out some transformations of word order, if these transformations are required and more precisely described with the translated word). Thus, as a result, the process of translation is from the most part *data driven* (i.e. input plus dictionary driven, to spell this out fully).

The dictionary of the project now contains (August 1988) about 7, 000 Czech lexemes, with all the information necessary for MT, as described earlier. Apart from words of common usage the dictionary covers terms from the field of computer operating systems. (The whole project is intended to translate operating systems manuals.) The dictionary is divided, as usual (cf. TAUM-73 or Kirschner 1987), into two parts. The first part is the dictionary of non-inflected words (adverbs, prepositions, conjunctions, abbreviations and the like) and words with irregular inflection: in this part, the heading of an entry is the word itself. The second (and by far the larger) part of the dictionary covers regularly inflected words; since this part does not stand alone in the system but is incorporated in the morphological analysis module as one of its main parts, the headings of entries are word stems (rather than, e.g., canonical forms such as nominative, infinitive etc.).

The Contents and Utilisation of Dictionary Entries

For the purposes of MT based on the approach described, the most different kinds of dictionary information are needed.

First, any Czech lexeme must be associated with (all) its Russian equivalent(s). Further, the dictionary must contain information on (both Czech and Russian) part of speech and inflection class, for the purposes of morphological processing. Naturally, such morphological information alone would be insufficient for high quality translation: thus further information concerning morphology is added, this time of a contrastive type (e.g. the information with the relative pronoun "kdo/kto" ("who"), the number of which might be arbitrary in Czech but only singular in Russian - cf. the Czech construction "všichni (=pl), kdo bojují (=pl) za mír" with the

Russian one "vse (=pl), kto boretsja (=sg) za mir" ("all, who fight/fights for Peace")). Another piece of information contained in the dictionary concerns syntax. The canonical example of dictionary-based syntactic information are the valency frames (case frames) associated with lexical units. One item of a valency frame (i.e., slot to be filled by some word) consists of three parts: the Czech morphological form (typically, simple or prepositional case) of the word required by this item, its Russian morphological equivalent (which might be, of course, quite different, e.g. for verbs with different government) and information about the semantic nature of the required word. At this point it should be mentioned that closely connected with frames are also lexical redundancy rules, operating on them after morphological but before syntactic analysis, creating passive parts of frames of transitive verbs, changing accusative case to genitive of direct objects with negated verbs (in the Russian part) and making some other minor changes. But there is also other syntactic information resident, the most important of which is probably again that of a contrastive type, which enables the system to deal with syntactic differences between Czech and Russian (e.g., different word order: this phenomenon occurs particularly often in cases where Russian equivalents are of different part of speech than their Czech counterparts. A notorious example is that of attributes of nouns: Czech adjectives must be quite often translated as Russian genitive noun attributes, e.g. Czech "stavové slovo" and Russian "slovo sostojanija" ("(program) status word" / "word of status")). Further information in the lexicon concerns semantics; this is intended, however, just as an auxiliary means of support for syntactic analysis and/or translation and, hence, is quite simple, amounting just to associating bundles of semantics makers such as Concrete, Human, Time, Software-Product etc. with nouns, advebs and some adjectives. During the parsing process these bundles, together with the semantic requirements in frames, help to make the "right" (in the sense of "more probable") choice between competing alternatives of syntactic analysis or translation (e.g. the Czech adjective "vstupní" ("input") should be translated as either "vchodnoj" ("entry"), with concrete nouns, or as "vvoda" ("of input") with abstracts). Particular problems occur if a one-word Czech lexeme has to be translated as more than one Russian word, especially if the sequence of Russian equivalents can be discontinuous: this happens quite often with verbs (e.g. Czech "zkompilovat" ("to compile") and Russian "osuščestvit' kompiljaciju" ("to perform compilation"), cf. also the sentence "Tuto proceduru programátoři zkompilovali včera." (Czech) with "Etoj procedury programmisty *osuščestvili kompiljaciju* včera." (bad Russian) and "*Kompiljaciju* etoj procedury programmisty *osuščestvili* včera." (correct Russian)). Again, all information necessary for this purpose is stored with the lexical entry in the dictionary. A problem all of its own in MT is always thrown up by terminology, especially by compound terms; in this project, all the information needed for solution of this task has been shifted to the dictionary, using all the means of dictionary information mentioned thus far plus a network of specialized "semantic" features, one for each compound term. These special features serve then for binding together the individual words of the term (as well as for preventing some other word not belonging to the term from intervening) in the process of analysis, and this in exactly the same way that the regular semantic features do for the choice of pragmatically plausible parses. In this approach, no special term-oriented extensions of the syntactic analysis and synthesis modules were needed, which was important for the efficiency of the system.

Advantages of the Approach

The adjustments made to the classical transfer MT philosophy, as described in this paper, yield the following practical advantages: higher efficiency of the system (in both CPU-time and memory requirements), improved "robustness" of the system (just one "dangerous" interface, between Czech analysis and Russian synthesis, instead of two in the classical transfer approach), better modularity (more information is stored in separated dictionary entries rather than in more or less monolithic programs of analysis and synthesis), thus making it more easy to perform debugging and extend the system.

References

Hutchins W. J. 1986. *Machine Translation: Past, Present, Future.*
Chichester: Ellis Horwood Ltd. publishers.

Kirschner Z. 1987. APAC 3—2: An English-to-Czech Machine Translation System, Explizite Beschreibung der Sprache und Automatische Textbearbeitung vol. XIII, MFF UK, Prague.

Slocum J. 1985. 'A Survey of Machine Translation: its History, Current Status and Future Prospects' in *Computational Linguistics* 11: No.1, January-March.

TAUM—73. 1973. Report, University of Montreal.