Czech-to-Russian Transducing Dictionary

Alla Bémová, Vladislav Kuboň
KAM - linguistics
Charles University
Malostranské nám. 25
CS-118 00 Prague 1

0. A bottleneck of all productional-oriented machine translation systems is to handle those words which are not included into main dictionaries of the particular system. In this paper we want to describe one possible approach to this problem, based on the idea of the so-called transductional dictionary. The first part of the paper is devoted to the results of our empirical inquiry into the problem in closely related languages, the second one to the description of the implementation of the transducing dictionary in the Czech-to-Russian machine translation system RUSLAN (Hajič, 1987).

1.1. The basic idea of the transducing dictionary, originally developed by Z.Kirschner for the English-to-Czech MT system APAC (Kirschner, 1987, Hajičová and Kirschner, 1987), is quite simple. Among the words which failed to be found in the basic dictionaries of the system, are some special sets of words, which are very similar in both languages. The goal of the transducing dictionary is to translate those words properly according to the rules, dealing with regularities between source and target shape of them. This way of handling lexicon as an open system differs substantically from the traditional one. It is very useful especially in the translation of the scientific or technical texts where many international words are used.

The transducing dictionary (TD), originally used for MT between English and Czech languages, handled mostly international (Greek-Latin) words. We have assumed that there will be a greater number of lexico-derivational parallels between two closely related languages and that they will contribute to a successful analysis to a greater extent. This assumption turned to be false.

1.2. Let us now turn to some issues of a contrastive analysis of Czech and Russian to investigate, where the idea of TD could be applied for the translation between these two languages.

For instance, the Czech ending -ace is translated as -acija, as in modifikace - modifikacija "modification" etc.; thus, e.g. if the Cz. word "emulace" is not found in the main dictionary, it is translated by the transducing rule as "emulacija". A word translated in this algorithmic way may be subdued to the procedure of orthographic changes and modifications, in which some rules of Cz.-R. alternation of graphemes are applied, as, e.g., ě→e, h→g, y→i, x→ks, au→av, ou→u, the initial e→e, etc.

It goes without saying that such a transduction is accompanied by a risk of creating errors or is impossible, e.g., in case that the output language lacks such a direct derivational variant. However, even in such cases it is possible to apply along with the transducing dictionary further "emergency rules" and to determine, on the basis of the final segment of the word, at least the type of the word and its grammatical characteristics, and thus to prevent the termination of the analysis of the whole sentence due to a single non-identified word. Of course, every irregular lexical correspondence can be included in the main dictionary.

1.3. The rules of TD, as indicated above, can be used with several degrees of complexity: from the possibility to translate "mechanically" without any changes or only with some orthographic modifications as in elektroskop - elektroskop - "electroscope", through the necessity of certain modifications of the word-formative segment or ending according to the needs of the target language as in algoritmus - algoritm - "algorithm" to the limitation to a mere identification of the word as for its word class, or, as the case may be, its morphological or semantic characteristics as, e.g., with tlačítko - knopka "push-button", where the ending segment -tko points to the noun, neuter, denoting an instrument (means).

According to these degrees, the set of derivational morphemes productive for the domain of computer techniques can be classified into several groups; not always we are, of course, concerned with a suffix in a strictly linguistic specification of the term; rather, we deal with a segment which has for the given pair of languages the required properties (it may consist of two suffixes or their parts, as, e.g., in -kce, -uce, and also of an original word stem, as gram; we quote here segments in the form of nominative sing., i.e. together with the inflectional ending).

1.3.1. The most substantial part of the first group consists in words denoting the names of machines, devices and instruments, and also some names of agents of events, of some properties and processes. These words can be translated to Russian by a mere transcription. The following suffixes are members of this group:
-skop, -metr, -tor, -graf, -tron, -fon, -at, -log, -ant, -ent, -tura, -ika.

1.3.2. The next group contains nouns and adjectives the form of which must be modified according to models customary in the target language, which involves some minor changes in the derivational segment or ending. The respective suffixes are:
-smus/-zm, -ace/-acija, -kce/-kcija, -uce/-ucija, -ance/ -ancija, -ence/-encija, -ie/-ija, -gram/-gramma, -aze/-azis, -eze/-ezis, -ita/-ost', -ium/-ij, -ista/-ist.

Certain regularities of translation of derivational suffixes can be observed also with adjectives, which form an important part of terminology, cf. the following correspondence:
-ický/-ičeskij, -ární/-arnyj, -ální/-alnyj, -antní/-antnyj, -entní/-entnyj, -ční/-cionnyj, -ivní/-ivnyj.

1.3.3. The third group includes semantically uniform and productive classes of words, as, e.g., deverbative nouns ending with -aní, -ení, which denote in Czech a process or its result (kódování - encoding, spouštění - switching on), nouns in -ost denoting a property (prašnost - dustinnes, vlhkost - humidity), nouns in -tel denoting an animate or inanimate actor (odběratel - consumer, ukazatel - marker) and also nouns in -ště with the meaning of a certain place or space (naleziště - deposit, pracoviště - working place) and nouns in -tví with the meaning of a feature, workshop or property (množství - a number of, pekařství - bakery, bohatství - richness). All these word-formative types have a corresponding equivalence in Russian, with minor modifications. However, in the course of the long-term development, the semantic shifts of the word basis prevent the possibility of the translation of these types only by means of the word-formation correspondence of the transducing dictionary. TD cannot do here more than specify the word class or the gender. These data can be used in the syntactic analysis of the source language but in the R. output the word has to be marked as not beeing found in the dictionary.

1.3.4. There is one more group of words we encounter in technical texts which is productive in Czech and has a clearly specified meaning. There belong the nouns in -č (-ič, -ač, -eč): chladič - cooler; these nouns denote first of all a device or instrument, sometimes also an animate actor. The suffix has no word formation parallel in R. and is substituted by other suffixes, or frequently by a complex naming unit: překladač - transljator "translator" or přepínač - pereključatel' "switch". A similar situation occurs with the suffixes -átko, -ítko, -dlo, -árna, -ovna. Also there suffixes have no corresponding equivalent in R.

In spite of the fact that in this group, the derivational means determine classes characteristic both as for their meaning and form, this fact is made use of onl; partially - the transducing device easily and correctly identifies their word class and can assign to them morphological, or as the case may be, semantic characteristic.

3.1. The implementation of the transducing dictionary was to a great extent influenced by the programming language used, namely the systems Q (Colmeraurer, without date) in which the whole system RUSLAN is written. Systems Q provide means for working with grammar directly in the form of rewriting rules, which are realized by means of a labelled graph; the use of a rule means adding a new edge. In order to limit the number of rules, the programme (i.e. the set of rewriting rules) is divided into phases. At the end of each phase , the redundant edges are removed. The whole system RUSLAN consists of 16 phases.

The TD does not immediately follow the main dictionary (i.e. the dictionary of

stems) because the mechanism of morphological analysis cannot distinguish at that point, which words have been found in the dictionary and which have not. This is due to the fact that the identification is given by non-empty intersection of the pattern of the stem and the patterns that come into consideration as the patterns of the ending; this intersection is identified by the phase immediately following the dictionary look-up. In the next phase, a preliminary syntactic analysis is carried out, in the course of which the unambiguous words found in the dictionary and immediately depending on each other are put together. At this point, the rules of the TD are applied.

The forms which have not been found in the main dictionary pass through the phases unchanged. In the first phase of the TD two kinds of rules are applied - the first one dealing with the groups from 1.3.1. and 1.3.2. and the second one with the rest (1.3.3. and 1.3.4.). The only difference among those two types of rewriting rules is that the second one does not transcribe the Cz. lexical value of the word into the R. Both add morphological characteristics and semantical features to particular endings.

The first phase of the TD contains also the first portion of rules rewriting the result of the application of the above mentioned rules (of the first type). The following transcriptions are carried out : hi→i, x→ks, hev (at the beginning of the word) →ev, ou→u, hy→gi, all long vowels are changed to short, ř→r, ě→e, ď→d6, ť→t6, ň→n6 (6 is a code for the R. soft marker), gy→gi, ky→ki. In this way, the original Czech stems (to which the R. endings are attached at this point) are transcribed into the corresponding R. forms.

The second phase of TD is added in RUSLAN to the phase of Czech analysis in which the disintegrated stems are put together again. This phase of TD has two main tasks:

(a) to finish the transcription of the Czech stems into the Russian ones, which for technical reasons could not be done in the preceding phase,

(b) to modify the form of representation in such a way as to be consistent with the words that have been found in the dictionary; the rules remove the auxiliary symbols and put together disintegrated Russian lexical unit.

The third phase of TD processes a rather small set of words which have not been found in the main dictionary and for which none of the rules of the preceding phases of TD can be applied. This unrecognized words are great obstacle for the build-up of the dependency tree because they do not allow the parts that stand to their left and to their right to be put together in the course of the syntactico-semantic analysis. As a temporary solution of this problem, the rules of this phase rewrite all hitherto unidentified words as nouns in neuter sing. For a more reliable solution, the work is in progress to distinguish between the endings in a more subtle way (e.g. those forms ending in -ě should be assigned the word class of adverbs, etc.). The default procedure of the third phase of TD arrives at a dead end in case the unidentified word is a verb, since then the sentence lacks its key element - the root of the tree.

It is evident that none of the guiding lines applied in the TS in its present form is 100% valid. However, it has a great value for the experimental testing of the system because it filters out words that have to be added to the dictionaries of the system.

REFERENCES

Colmerauer A.: "Les systemes Q ou un formalisme pour analyser et synthetiser des phrases sur ordinateur". Mimeo, without date. Montreal.

Hajič J.: "Ruslan - an MT System Between Closely Related Languages, In: Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics, Copenhagen 1987, p.104 - 108.

Hajičová E. - Kirschner Z.: "Fail-Soft ("Emergency") Measures in a Production-Oriented MT System", In: Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics, Copenhagen 1987.

Kirschner Z.: "APAC3-2: An English-to-Czech Machine Translation System". Explizite Beschreibung der Sprache und automatische Textbearbeitung XIII, Prague 1987.