COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

ON A DEVICE IN DICTIONARY OPERATIONS IN MACHINE TRANSLATION

Zdeněk Kirschner

Department of Applied Mathematics
Faculty of Mathematics and Physics
Charles University
Prague
Czechoslovakia

A special programme converting classes of words of international usage directly from English to Czech is described in its application in an experiment of machine translation as well as in general environments. The words undergo special morphemic analysis, they are adapted morphemically and orthographically to the target language form and,in the experimental version, they are assigned pertinent grammatical and semantic information.

Dictionary operations in automatic processing of natural language involve, as a rule, a number of problems, most of which being connected with the fact that lexicons generally tend to grow beyond tolerable measure. That is also why the ideal but hardly feasible principle of listing in the lexicon only "irregularities" has retained its methodological appeal, especially in the automatic analysis of natural language. Various solutions may be found to the problem of reducing the size of dictionaries; the one presented here may be of some interest, because the algorithm described performs the task of substituting classes of target language equivalents for classes of source language expressions.

Replacing the source language words by target language equivalents must be regarded as an operation involving changes "irregular" if not by nature with some pairs of languages, then certainly in point of fact in actual practice everywhere. Most languages that have or are supposed to have had a common ancestor differ at the lexical level to such an extent that it is impossible to cover the etymological changes by any applicable set of rules, and even if this were possible, semantic shift would interfere in an incontrollable way. However, the impact of some relatively recent historical changes (economic, political and cultural)has led to a specific and partial rapprochement of languages: the assimilation of words of foreign origin caused by these processes usually happens in conformity with standard morphemic and orthographic rules and, since most of the new acquisitions stay beyond the reach of common usage, semantic shift is rare.

On this basis the so-called "translational" or "transducing" dictionary (henceforth TD) has been developed in the experiments with machine translation carried out in the linguistic group at the faculty of mathematics and physics, Charles University, Prague. This

device converts the words of international usage from their source language (English) form into the morphemically and orthographically adapted target language (Czech) counterparts in a three-step procedure, which involves also the modifications in the corresponding derived parts-of-speech, i.e. adjectives derived from nouns, adverbs from adjectives, etc. In its present form, the TD covers the classes of English nouns ending in CE, CY, EGY, ENT, ER, ERE, GRAM, GRAPH, ICS, ID, ISM, ITY, IUM, ODE, OGY, ON, ONY, OPY, OR, ORY, PHONE, PHY, SCOPE, SIS, TRY, URE, adjectives in ABLE, IBLE, AL, ARY, IC, IVE, OUS, RSE, and verbs in ATE, FY, ZE, and it can be extended as required. Most of these classes fall into more subclasses to meet the specific requirements based on grammatical, semantic, or, as the case may be, orthographic distinctions. The equivalent forms need not differ at all, as e.g. in IMPEDANCE (but the adjective derived is IMPEDANČNÍ), or they differ very little (DIODE ≒ DIODA, GRAPH =GRAF, etc.), or differ considerably (OPTIMIZATION = OPTIMALIZACE, INTENSIFIER = INTENZIFIKÁTOR, DEMONSTRATE = DEMONSTROVAT, ADAPTABLE = ADAPTOVATELNÝ).

In the overall algorithm, the main dictionary precedes the TD, so that all grammatically and/or semantically idiosyncratic words formally belonging to the classes enumerated above can evade the treatment in TD: e.g. ICE, ICY, WENT, SCOPE, or even APPLICATION (a noun with partially idiomatic properties taken over from the corresponding verbal frame), etc. The forms of words that passed through the main dictionary, failed to be identified, and were reconstructed again, enter the first step of the TD procedure: a special morphemic analysis concerning forms of unrecognized words that are potential members of the classes covered by TD. Thus, e.g., the normal form ILLUSTRATE,(+ED) is reconstructed from the form ILLUSTRATED, STRATEGY (+S) from STRATEGIES (but not OPTIC (+S) from OPTICS, where a restriction interferes). Some words undergo special modifications in the first step, mostly to be adapted to more frequent usage or to more profound changes in the target language (e.g., with both kinds of adaptations in one, MINIMISE becomes temporarily MINIM-ALIZE; LOCALIZE, of course, remains unaffected).

The second step selects the words to be treated and replaces the source language suffixes or endings by the corresponding target language ones: as examples, three rules can be quoted, the first for nouns ending in ITY (permeability, viscosity, thermoelasticity, etc.), the second for verbs in ATE (demonstrate, illustrate, etc.), and the third for adjectives in SSIVE (aggressive, passive, massive, etc.). The words enter this step desintegrated, as lists of ordered characters. In the following rules, $\underline{a}$, $\underline{b}$ are variables for labels - individual characters or strings of characters - $\underline{u}$, $\underline{v}$, $\underline{w}$ variables for lists - e.g., one or more characters, labels, etc.; they can be empty lists, too. Capital letters and digits constitute constants; the double sign ⨯⨯ or ⊠⊠ indicates that the following label is desintegrated and represents a list of characters. The rest of characters used convey grammatical, semantic and/or technical information. The conventional signs $\wedge$, $=$, $\neq$, $\cap$, $\subset$ have their usual meaning, viz. conjunction, equality, non-equality, set-intersection and set-inclusion, respectively; == means "rewrite as".

```
ATOM (u,I,T,Y) == PN (u,I,T,*(Ø93V2),*A,*PROP,/,Ø).
ATOM (u,A,T,E) == PV (u,U,J,*(21Ø2WW5U1X3),1(*A,*H,*C),2(*A,*C,*OB),
                                        /,Ø).
ATOM (u,a,⊠⊠b) == PA (u,a,⊠⊠SI2VNI2,*(Y),*A,*C,*MNR)/a↦S\ b↦SIVE.
```

The third step is closely connected with the second and performs the necessary orthographic adjustments. E.g., the rule

CØ(u,⊠,a,b,v,#,w) == CØ(u,a,⊠,v,#,w)/a=T∧b=H∨a=b∧({b}∩{C,O,S})=Ø.

applied recursively until v equals one single character or is empty, changes TH into T; any doubled character (with the exception of C,O,S) becomes single. Very similar rules handle the groups PH, QU, etc.; e.g., another rule

CØ(u,⊠,a,b,v,#,w) == CØ(u,K,b,⊠,v,#w)/a=C∧{b}⊂{A,L,O,R,T,U}.

applied recursively in the same way changes C which immediately precedes A or L or O or R or T or U into K. There is a set of such rules dealing with groups of two or, as the case may be, more characters; complementary rules control the movement of the ⊠ sign in case the first character, the second or the following ones do not correspond to the condition. To quote the simplest example, the following rule moves the ⊠ sign by one place only in case the first character (standing in the position indicated by the variable a) is not T or P or Q or C, and the two characters tested are not equal:

CØ(u,⊠,a,b,v,#,w) == CØ(u,a,⊠,b,v,#,w)/({a}∩{T,P,Q,C})=Ø∧a≠b.

The actual writing of rules is, of course, more economical, e.g. the left members are not repeated, etc. Thus, original English words like THERMOELASTICITY, PHILOSOPHY, QUANTIFIER, ILLUSTRATE, MASSIVE become Czech stems TERMOELASTICIT, FILOSOFI, KVANTIFIKA2TOR, ILUSTRA2TOR, MASI2VNI2 (note that the digit 2 following a vowel denotes the diacritical mark "´" indicating length), respectively; first, of course, they are provided with all necessary information, reintegrated and brought to the canonical form required by the system.

In our experimental system, the TD operates on a relatively narrow, semantically circumscribed domain of scientific and technical sublanguage concerned with microelectronics: this makes it possible to keep the semantic description apparatus at an adequate and, at the same time, sufficiently general level, a condition on which the "profitableness" of TD depends under circumstances described. The programme of TD written in Colmerauer´s Q-systems (Q-language) consists of less than 90 instructions (rules with a relatively high degree of recursion involved) and it is supposed to cover a set of thousands of lexical items, most of which represent very frequent terms or term components. In the described environments (experimenting with English-Czech machine translation, microelectronics) the device disclosed has proved extremely effective and there are reasons to believe that it can serve its purpose in other conditions as well: e.g. in application to other pairs or multiplets of languages, in other domains, and/or applied in a different, theoretically more scrupulous manner, it may give even better results.

The latter observation deserves some comment. So far, the application of transducing dictionary in the framework of our experiment has benn considered. However, this particular way of application must not be regarded as anything else than purely experimental testing of this device in partially simpified conditions; in fact, the results of such a testing have been reported on in the preceding paragraphs. In a full-fledged system of natural language analysis and synthesis, the considerably more developed semantic apparatus would impose much more rigorous restrictions on the constitution of uniform, semantically homogeneous, classes to be covered by TD, which might impair its effectivity considerably. The proper place of TD is at the morphemic or morphonemic level. If the operation of TD is confined to these domains only (grammatical and semantic information being assigned in another place), its use can be generalized, so that, in substance, all words formally belonging to the classes defined purely on morphemic and orthographic grounds can be transduced by it. Exceptions are rare; some of them can be solved in the framework of special morphemic de-

composition (as in the above example - MINIMISE vs. LOCALISE), some must be handled by the dictionaries that precede the application of TD in the algorithm - statistics must decide what should be regarded as exceptional (e.g. the ending -XIAL in English may give -XNÍ or -XIÁLNÍ in Czech - EPITAXIAL vs. COAXIAL ; however, it should be observed in this place that both versions - the generally accepted one (EPITAXNÍ) and the other (EPITAXIÁLNÍ) would be undoubtedly understood without any difficulty by the reader). In this connection, it should be pointed out, that, in such highly specialized domains as microelectronics or computer system programming,etc. the widely used practice of taking over and adapting foreign expressions affects in a high degree regions traditionally "protected" by normative grammarians and patriotic terminologists (GETTER = "GETR", CHIP = "ČIP", HOLD ="HOLDOVAT", ABEND = "ABENDOVAT", PRINTER = "PRINTR", etc.); in our opinion, this practice is very often fully justified and it should be rather coordinated than condemned, since it makes prompt and correct understanding between experts possible, and it undoubtably contains rational and progressive tendencies contributing to a new phase in the rapprochement mentioned above. Whether a new or adapted TD like device may be found useful in this domain, too, remains to be seen.