FAIL-SOFT ("EMERGENCY") MEASURES IN A PRODUCTION-ORIENTED MT SYSTEM

Eva Hajičová and Zdeněk Kirschner
Fakulty of Mathematics and Physics,
Charles University
Malostranské n.25, 118 00 Praha 1, Czechoslovakia

ABSTRACT

A system of fail-soft (emergency) measures for a production-oriented MT system is discussed, stating first the specific purposes of such a system, and showing then, how these measures are being used in the system of English-to-Czech machine translation as prepared by the group of mathematical linguistics at Charles University in Prague.

1. In view of a production-oriented system of machine translation, under the present-day conditions, one should keep in mind that the end-user expects to have at his disposal a complete text rather than an alternating sequence of (sentence) segments and blanks. On the other hand, everyone who has ever made even a perfunctory look at the problems involved in MT would agree that there is no such thing as a "complete" MT system, neither in the dictionary nor in the grammar part of it. Also, as is commonly accepted nowadays, any system with texts written in natural language at the input should provide measures for some kind of treatment of ill-formed input. Thus it is inevitable to consider, first, what type and quality of translation can meet the demands of a prospective user and what kind of translation is realizable under the given conditions, and, second, to decide what is to be sacrificed from, and what added to the system to make it work in a production enviroment.

In the present paper we would like to outline one aspect of the approach taken up at the start of our experiment APAC3-2 - the English-...-Czech machine translation system for translating INSPEC abstracts from the field of microelectronics (for a description of the history of the MT efforts of our team, see Hajičová, 1986; the APAC series is described in full detail in Kirschner, 1982; 1984; in press). We should emphasize that in the conditions within our reach, we aim at a satisfactorily accurate rendering in target language of the contents of a relatively simple text in source language, which would suffice for such a system to be applicable in information acquisition and would be capable to meet the main requirements set up by average users.

2.1 The specific purposes of the system of fail-soft ("emergency") measures to overcome anomalous input phenomena and partial failures of the MT system can be stated as follows:

- The whole processing is divided into several stages to identify and treat more probable interpretations of some structures preferentially - to make the "preferential" approach (see below, Sect.2.2) a reality.

The drawback is that in these circumstances the danger of a dilemmatic situation resulting in a cul-de-sac impasse increases. A special device has been introduced to overcome such an abnormal end: instead of eliminating such a defective string, the application of the phase in question is suspended - the program processing this string skips the phase and continues in the next one. The rules that compensate for the lacuna can be either the rules that in the framework of "preferential" approach take up the role of their more strict predecessors, or rules added particularly for this purpose - to deal with the most undesirable consequences of such an omission.

- In the analysis, the emergency rules interpret unrecognized elements and integrate them into more complex structures.

- In the synthesis, they help to produce an output that makes sense, corresponds to the source language, and is easier to post-edit.

- Whenever it is possible, they attempt at forming target language equivalents for the unidentified elements, either by adapting international words or by "czechizing" English dictionary forms by enduing them with qualities and forms proper to their presumptive Czech counterparts - e.g., gender, suffixes, etc.

- With some classes of words, they serve as general dictionary rules provided the sets of semantic features, frame information and other necessary outfit of individual members of these classes correspond to the standard apparatus assigned to their representation in the framework of a general device, and their orthography ensures forming correct equivalents in Czech.

2.2 The fail-soft measures can be characterized as consisting of three main parts: the first two concern elements not found in the basic dictionaries and the third concerns failures to arrive at an accomplished parse. (We leave aside a discussion of the unification of orthography - such as American and British usage, different ways of spelling, use of hyphens etc. - which comes before the first device described here.) In a sense, there is a set of other rules of "emergency" character: general rules (which can be called "sweeping rules") designed to operate after all more specific rules failed to apply - e.g., in the formation of compounds or in nominal syntax in general, etc.; however, this being a constitutional component of what we call "preferential" approach, we shall confine ourselves to describing only the former three sets. To avoid a possible misunderstanding, we should make clear that when we call our approach "preferential", it is only the name that it has in common with Wilks' "preferential semantics". In our system, we apply a rather trivial and simple principle with the aid of which the different probability of interpretation(s) of some parts of a string is taken into account and exploited. The most probable solutions are covered by the rules first and with as detailed an accuracy as possible; the next probable solution is offered in some of the subsequent phases, etc., under more liberal conditions. The "sweeping rules" come last. That is also the reason why we write "preferential" with quotation marks.

2.21 The first device aimed at intercepting and interpreting words that failed to be found in the basic dictionaries is the co-called transducing dictionary (TD). Its task is to interpret the still unrecognized words according to their typical and (mostly) productive suffixes (the inflectional endings being detached and dictionary forms reconstructed by morphemic analysis in the preceding steps), and to

assign to them part-of-speech and semantic information. Thus, e.g., words ending in -ER, -OR, -GRAPH, -ODE and some others are interpreted as nouns, concrete, instruments, capable of being substituted for human actor; words ending in -CS, -CY, -ESS, -TUDE are supposed to be nouns, abstract, properties and, as distict from those ending in -ITY, -ICS, -SM, -SHIP, -HOOD, -THM, which otherwise have the same semantic characteristics, they form adjectives in a regular manner in Czech; the endings -FY, -ATE, -ISE (-IZE), -DUCE indicate verbs that can be both transitive and intransitive, of causative and (semi) terminological character,yet not allowed to form adjectices of the purposive type. A number of adjectival suffixes is contained, too, viz. -ARY, -AL, -RSE, -IVE, -OUS, -IC, -BLE, -LESS, -ANAR, -LEAR, -NEAR, -OLAR, -ULAR. In all, about 50 classes of nouns, 13 of adjectives and 4 of verbs are covered by the TD device.

Two further pieces of information should be added, the first being probably superfluous: 1) All words having such suffixes but different properties as regards their part-of-speech category, semantic features, etc., are supposed to be contained in the basic dictionaries. 2) Most of the classes of words treated by the TD are international words of Latin or Greek origin; they can easily be "transduced" to Czech by relatively simple procedures; some of these procedures precede the TD operation as a part of a special morphemic analysis, but most of them operate in the synthesis, as an accessory to the English - Czech dictionary. A set of recursively applied rules (in several cycles) takes over the words identified by TD, desintegrates them, replaces the English suffixes by the corresponding Czech ones, and scans the bases for spelling configurations to be transformed or adapted to Czech orthography (replacing, e.g. PH by F, TH by T, C preceding A,L,O,R,T,U by K; S preceded by A, E,I,N,O,R,Y and followed by A,E,I,O is replaced by Z, etc.). Thus, e.g. PHOTOLITOGRAPHIC changes into FOTOLITOGRAFICKE2, CYCLOTRON gives CYKLOTRON, ISOSMOTIC is transcribed as IZOSMOTICKE2. To give an example of solving similar problems, let us consider the word ISOSEISMIC: to preclude the second S situated at a morphemic juncture from becoming a Z, would require either a special entry in the main dictionary - as one word or as combination of the prefixal ISO + SEISMIC, in which case the adjective must be contained in the dictionary - or some similar preliminary treatment in the special morphemic analysis preceding the TD; the latter way of treatment would probably represent the best solution, which may be generalized for all or most of the typical terminological prefixes involving analogous problems as IZOSEISMICKE2 - e.g., A-, INFRA-, PRE-, PERI-, SEMI-, SYN-, MESO-, MONO-, HYPER-, POLY- etc. (needless to add that this time it would be such words as ISOSMOTIC that would require a specific treatment, e.g. to process only SMOTIC - - from ISO + SMOTIC - in the dictionary). It should be remarked that, in principle, this part of the transducing device - orthographical changes - need not be separated from the front part operating in the analysis.

2.22 Words that remain unaccounted for after passing the TD phases - i.e., not found in the dictionaries and not belonging to any of the classes dealt with in the transducing device - are subjected to further analysis; those having typical verbal inflectional endings (-ING, -ED) are regarded as verbs, those ending in -LY are taken for adverbs provided that more than 2 characters precede and their tentative status is syntactically corroborated. The

rest are first treated as proper names and if the subsequent analysis fails to confirm this conjecture - i.e., they are not integrated into wider nominal complexes, e.g., as an apposition - they become nouns (which, by the way, happens to the tentative adverbs, too). The words identified in this tentative manner are "czechized", which in some cases might result in quite acceptable formations - e.g., if the original words can be taken an "international" or technically and terminologically univocal terms: GETTERING → GETEROVA2NI2, ABEND → ABENDOVAT - in other cases in more or less comical "macaronic" creations. In conclusion, it should be added that the original more ambitious idea of assigning to each unrecognized word (that does not carry any characteristic clue making the guess easier) three parallel tentative interpretations to let the syntactic analysis decide -noun, verb, adverb - had to be abandoned for reasons similar to those that led to the resignation in the case of hypersentential context. Too many possibilities, often combined with other parallel solutions, led to combinatorial explosion that (though often not assuming the character of an infinite loop) expanded the structures to such an extent that sooner or later an overflow became inevitable. So far, there is no remedy for overflow in our system.

2.23 The last, relatively simple, measure concerns cases where a single parse (or more parallel single parses) - i.e., trees covering individual input strings - failed to be formed in the last phase of the analysis; usually two or more partial trees are formed instead, which fact may be caused by anomalous structure of the input string, or owing to some partial failure in analyzing one or more substrings (e.g., when some element(s) or structure(s) were misinterpreted), or as a result of some subjective shortcomings in the program - - omission, error, etc. The synthesis program is able to process even such partial and fragmentary results and attempt at compiling an acceptable output, only a special character (@ or }) is placed in front of such output strings to signalize that they had been formed on the basis of defective results of the analysis. If necessary, a set of rules of a more or less ad-hoc character deprives "underdone" (sub)trees of all auxiliary structures (category labels, parentheses, separators, features, etc.) leaving only lexical values, and performs thus the finishing touches to bring the substitute output as close to readable and acceptable results as possible.

3. The outputs of individual phases can be obtained in the listing. Some of these phases, esp. the last-but-one phase fixing the state of things before the syntactic measures have been applied, usually preserve information enough to recognize and examine the unretouched results and to divulge the diagnosis of errors or shortcomings necessary for further progress. This is to say that most of the "emergency" devices operate at moments and in a manner which permit to examine the previous state of things, so that their action does not obscure the regular course of the processing and allows normal control of it. It should be added that a part of emergency devices has a temporary character dealing with omissions and bugs proper to the system under development. We are sure that at least some of them will become superfluons.

REFERENCES

Hajičová, E. (1986) Machine Translation Research in Czechoslovakia, Proceedings of the Int.Conference on Translation Mechanization, August 20-22, 1986, Copenhagen

Kirschner, Z. (1982) A Dependency-Based Analysis of English for the Purpose of Machine Translation, Explizite Beschreibung der Sprache und automatische Textbearbeitung IX, Prague

Kirschner, Z. (1984) On a Dependency Analysis of English for Automatic Translation. In: Contributions to Functional Syntax, Semantics and Language Comprehension (ed.by P.Sgall), Prague, 335 - 358

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