TRADITIONAL MEANS IN MACHINE TRANSLATION

Zdeněk KIRSCHNER

Matematicko-fyzikální fakulta UK
118 00 Praha 1, Malostranské náměstí 25
Czechoslovakia

Abstract:

The chronic problems of machine translation cannot be solved in a fully automatic way. Human intervention is inevitable. The development of "traditional" means in connexion with advances of computer technology represent most substantial contribution to further progress in the field of machine translation. Some of the problems are illustrated using the example of the APAC32 project.

1. The hopes for a successful solution of the chronic problems of machine translation (MT) have long been set on two fruitful and mutually dependent prospects: research in artificial intelligence (AI) and advances in the computing technology. The importance of the latter contribution is beyond dispute. As regards the former domain, some reservations must be voiced.

1.1. It can be stated with some tolerance that the missing information required for automatic understanding (or desambiguation) of natural language (NL) is supposed to be supplied by a computer model of the knowledge corresponding to the universe of discourse. The context of the analysed message constitutes an important part of this universe. Therefore, an essential component of such a model must draw on the texts processed. Thus, irrespective of the contingent form, organisation, etc., of the whole, the model would at least partially depend on the results of the analysis for which it is supposed to provide necessary information. This means that circularity is imminent. Even if the almost inevitable occurrence of elements not covered by any device in the system is disregarded, it is obvious that the model can be neither complete nor consistent.

1.2. Since there will always remain threats of failure caused not by accidental factors but by the intrinsic inadequacy of any system of MT, human intervention is inevitable, and the ideal of "fully automatic high quality translation" (FAHQT) (which, we suspect, is no longer believed to be able to ever come true, anyway) is impossible. While not denying potential merits of the contribution of AI, the above discussion should suggest that the development of means called "traditional" is equally important for MT. An example of an approach based on such means is our experimental system APAC32.

2. If we refer to our system here, it is not to boast that we have achieved any extraordinary success or that we have long duly appraised the above conclusions and reacted on them in an original way, etc. It is only to illustrate our conviction that there is still a fairly wide and long path open ahead of us within the confines of the traditional means. To say the truth, it has been our material situation that forced us to rely exclusively on them and to dispense with anything more sophisticated. This had to be said to clear us of a suspicion that we are making a virtue of necessity.

2.1. APAC32 is a descendant of the Montreal TAUM series. It has been implemented on computers of the type of IBM 370 to translate into Czech English abstracts in microelectronics and, later, pumping machinery. Using Colmerauer's Q-systems the main part of the program builds linearized rooted-tree-like structures, which stepwise identify and interpret elements or groups of elements of the input units stating their character and function, dependency relations and position in the sentential context. Strings with multiple interpretations which had not been eliminated are represented by parallel structures giving multiple parses in the final stage of the analysis, but not necessarily multiple translations at the final output.

Basic or fully accomplished structures, which resemble predicate calculus patterns, have a finite verb at their root and individual participants in dependent positions. The sense (direction to the left or right) of an oriented edge (an arrow) representing a dependency relation - an information pertaining to the mutual projective position of the incident nodes - as well as the function of a dependent participant are indicated in a way that simulates the marking of edges in a graph. The synthesis starts by disintegrating the structures that result from the analysis. At the output of this stage, relatively simple trees representing individual words appear, with all the information necessary for generating forms of the target language. This proceeds in steps in which occasionally additional target-language-specific information has to be derived to render the synthesized structures complete and acceptable. Such adjustments are usually connected with the operations of transfer: while the action of its general rules mostly coincides with the opening phases of the synthesis, the information concerning the particular changes is contained in the dictionaries to be exploited in the concluding parts of the program.

2.2. The absence of any accomplished model of the universe of discourse and the temporary abandonment (for technical reasons) of any device alowing the involvement of hypersentential context in the analysis have, of course, endowed the system with a typical probabilistic character. In this connexion, especially the tactics occasionally referred to as "preferential" must be mentioned: some rules are applied repeatedly in subsequent stages, each time with conditions less rigid. The combinatorial power of the Q-systems had to be reduced by introducing several stages - partial grammars - operating before the syntactic analysis proper. Thus, e.g., a (partial) analysis of nominal complexes precedes that of verbal structures. Therefore, a special device registers schematically the context of each element in the sentence.

2.3. In simulating some functions of a model of the universe of discourse, the system of dictionaries represents the most important tool.

2.3.1. The basic dictionary information in APAC32 is a complex which consists of two main parts: information concerning the source language and that pertaining to the target language. These structures can be separated; they have been put together whenever possible with respect to the efficiency of the system. The internal structure of both these parts is almost the same and can be briefly described as follows: categorial information, lexical value, paradigmatic information, pointers to parallel meanings, valency frame, combinatory frame (prepositional, phrasal, special-liaison, etc., patterns), terminological specifications, special syntactic information, semantic features. Extensive though this apparatus may be, it should be stated that there are still possibilities - and a need, of course - to add further data. For lack of space, let us confine ourselves to three points only.

2.3.1.1. The apparatus of semantic features consists of four classes of features: a) features concerning the text vs. metatext structure, b) general semantic features, c) domain specific features, and d) features concerning terminological status. The number of features is limited for reasons of which the most important is that excessive detailedness leads to unwanted rigidity. However, a number of potentially very useful candidate features can be added. Assigning weights to features might be a solution to this dilemma, especially in the framework of the "preferential" tactics.

2.3.1.2. Some classes of words have been further classified to highlight their intrinsic properties in the translation environment. E.g., a special classification of verbs makes it possible to solve, at least in part, the problems of aspect in Czech in relation to Eng-

lish verbal adjectives (-ED, -ING forms). Much more can be done in this direction. Unfortunately, this will imply extensive empirical work including excerption and, if possible, organization of a usage-panel-like inquiry.

2.3.1.3. As concerns combinatory frames, also more information will be added on the possibilities of adverbial modification of nouns. Some changes and additions to the present organisation and contents of the dictionary entries are considered with a view to structures suggested in the Melčuk-Apresyan's model "meaning - text".

2.3.2. A specific dictionary device has been introduced in the terminological section of the dictionary system. Special rules control, or rather, guide the analysis of terminological complexes, making it possible to decide frequent ambiguous structures (e.g., INTEGRATED CIRCUIT SYSTEM as ((INTEGRATED CIRCUIT) SYSTEM) rather than (INTEGRATED (CIRCUIT SYSTEM))). In this way partial quasi-model of the specific domain can be formed whose elements are capable of recursive application to new combinations.

2.3.3. Another dictionary device deals with unrecognized elements - the so-called transducing dictionary (TD). TD relies on derivational morphology, assigning categorial information, and, in some cases, semantic status and other information to words hitherto "unknown" to the system, on the basis of their endings (e.g., -ING, -ED, -ESS, -ITY, -ION, -LY, -WISE, -FY, etc.); for some of them even successful adaptation to the target language is possible. The remaining unrecognized elements are regarded as nouns: first as proper, then, if this fails to be confirmed, common. A more versatile practice is planned, which will take into consideration other possible interpretations as well.

2.3.4. TD, as well as some other devices and rules can be also regarded as special fail-soft measures, though another component called "emergency rules" is included which performs this function as a specialized set of rules designed to reconstruct, complete or integrate into a (would-be) meaningful whole those structures that failed to reach the stage of an accomplished parse. In some respects, the role of such measures is problematic in relation to human intervention. Our system offers possibilities to introduce a special diagnostic device to recognize and classify the symptoms of a failure, so that more than the present simple marking of "suspicious" or "underdone" outputs can be presented to aid the postedition.

2.4. Ambiguities are treated in the usual way. It should be pointed out that in the translation between the languages in question, the principles of agreement so widely applied in Czech unmercifully reduce the chances to get over some types of unsolved ambiguities in an "unperceptible", i.e., accidental, way. These principles, as a rule, obstinately insist upon rendering implicit information explicit. That is why in some cases structures with ambiguous reference are translated by equivalents equally ambiguous or vague. E.g., with some classes of verbs, (clausal) participial modification with ambiguous dependence is replaced by prepositional or other constructions without any direct dependence: e.g., USING → WITH USING, CAUSING → WHICH (referring to the whole of the preceding or pertinent clause) CAUSES, etc.

2.5. This concerns also contrastive ambiguities and other asymmetrical relations between the two languages. In this connexion, it should be pointed out that one of the criteria for the classification of English verbs is the classification of their Czech counterparts. Thus, e.g., the verb SUPPOSE must be assigned information that the construction SOMEONE IS SUPPOSED TO... must be transformed to IT IS SUPPOSED (ABOUT SOMEONE) THAT SOMEONE... to make it correspond to the structure acceptable in Czech. Similarly, constructions like SEAT SAT ON BY... must be transformed with the aid of corresponding relative clauses.

Much remains to be done for the domain of conversion. Its productive aspects pose serious problems.

3. To come back to the opening paragraphs: the advances of computer technology, while not offering ultimate solution of problems detrimental to the efforts to achieve the ideals of FAHQT, will undoubtably liberate the MT from the curse entailed by its usually more or less immediate subservience to various practical applications - the strict limitations of computer time and storage - which so often represented the only obstacles in introducing many a useful and, sometimes, even very necessary device, process or approach. Most of the prospective extensions, innovations and other changes require profound empirical examination and more linguistic field-work than, up to now, we were able to expend.

References

Kirschner, Z. (1982) A Dependency-Based Analysis of English for the Purpose of Machine Translation. Praha, Matematicko-fyzikální fakulta UK.

--- (1987) APAC3-2: An English-to-Czech Machine Translation System. Praha, Matematicko-fyzikální fakulta UK.