THE NATIONAL BUREAU OF STANDARDS METHOD OF SYNTACTIC INTEGRATION

Ida Rhodes
National Bureau of Standards

Those of us who are now engaged in efforts connected with MT are about at the stage where the early pioneers in mathematics were before the appearance of mathematical journals. Each group had its own definitions, symbols, methods, and techniques. Mathematicians were duplicating each other's efforts, or else wasting valuable time in trying to solve problems already proved unsolvable. The development of communication between them brought in its wake a charming and useful custom -- that of issuing friendly challenges to each other. These came in two types. In one type, a mathematician would disclose his newly-found method for solving a certain problem, and he would challenge his colleagues to produce examples for which his method would fail. In the other type, the mathematician would list the problems which baffled him, and he would challenge his friends to find solutions for them. Today I should like to emulate the ways of my mathematical ancestors by offering challenges of both types.

Let me first ask your forgiveness for the hurried manner which I shall sail through my remarks, under the lash of fugiting tempus. However, most of our colleagues have done us the honor of visiting our Bureau in Washington, and have already been exposed to as much orating on my part as their forebearance could stand. Moreover, we have recently mailed a copy of our latest report, A New Approach to the Mechanical Syntactic Analysis of Russian, to every person on our mailing list. In addition, it is a pleasure to report that the Harvard group has found our method valid, and that at least one member of its delegation is planning to treat some aspects of our scheme in detail. Before undertaking the task of MT, we investigated the types of difficulties likely to be encountered, and found that they could be classified under 10 headings. We have been able to cope, so far, with only the first five of these, which relate to syntactical integration. These five difficulties are listed below in order of increasing complexity.

(i) The stem of a source word is not listed in our glossary. This will occur quite often in our translation scheme, as we intend to omit from the glossary the majority of non-Slavic stems.

(ii) The target sentence requires the insertion of key English words, which are not needed for grammatical completeness of the source sentence. For instance, the complete Russian sentence ОН БЕДНЫЙ (literally: "He poor") must be translated as: "He (is) (a) poor (man)".

(iii) The source sentence contains well-known idiomatic expressions.

(iv) The occurrences of a source sentence do not appear in the conventional order. Sober writing, without color or emphasis, employs few inversions. Our method, which consists of predicting each occurrence on the basis of the preceding ones, works quite well in that case. But such orderliness cannot be expected to hold for long stretches of the text.

(v) The source sentence contains more than one clause.

We have culled a Russian sentence from Volume V of Chebyshev's "Collected Works", which illustrates all the five difficulties enumerated above.

ПОКАЗАТЬ БЕЗ ПОСРЕДСТВА ТРАНСЦЕНДЕНТНОГО
(To) demonstrate without (the) means (of) [transtzendent] -al
АНАЛИЗА ОСНОВНЫЕ ТЕОРЕМЫ ИСЧИСЛЕНИЯ
[analiz] (the) basic [teorem] –s (of) calcul/ation/us
ВЕРОЯТНОСТЕЙ И ВМЕСТЕ С ТЕМ ГЛАВНЫЕ
(of) probability-s and at the same time (the) main
ПРИЛОЖЕНИЯ ИХ КОТОРЫЕ СЛУЖАТ ОПОРОЮ
applications (of) them , which serve (by/as the) aid
МНОГИМ ЗНАНИЯМ , --ВОТ МЫСЛЬ
(to/for) copious knowledges , -- here (is) (the) idea ,
МНЕ ПРЕДЛОЖЕННАЯ ГОСПОДИНОМ С. Г.
suggested (to/for) me (by) Mr. [S.] [G.]
СТРОГАНОВЫМ
[Stroganov]

This sentence constitutes our first challenge. We list below the steps entailed in translating sentences of this type.

1. Store, externally, Sample Glossary
2. Store, internally, lists of
   a. Frequently used words
   b. Pseudo-Prefixes

c. Pseudo-Suffixes
d. Pseudo-Roots
e. True Endings

We wish to point out that the pseudo-elements given in our lists have no semantic significance (except in a few rare instances); they merely constitute frequently used combinations of characters which we utilize for the purpose of minimizing storage space and retrieval time. The scheme for carrying out the above two steps is by no means a final one, since the form of the final glossary will not be determined before the completion of our translation scheme. The far more suitable equipment which will be available at that time will also have great bearing upon the glossary construction. At present, each entry of the externally stored portion of the glossary contains a highly compacted version of the stem of a Russian word, accompanied by copious material exposing all pertinent aspects of the stem. In addition to the stem glossary, we have also stored, internally, a special list of complete words, exactly as they might occur in the source text. Being confident that automatic reading machines will be available to us by the time we are ready to put our scheme into production, we assign a distinct 6-bit number to each possible source symbol (except those occurring in mathematical formulae). The lists enumerated under step 2 are used in the following iterative scheme, which is carried out for each source occurrence.

3. Read-in Source Occurrence (SO)
4. Decompose SO, if necessary
5. Intersort Pseudo-Root, if any

By means of these lists, we separate the ending, each pseudo-prefix and pseudo-suffix, and the pseudo-root of the occurrence in question. The pseudo-root is intersorted with similar elements of the previous occurrences. When the sorting file is filled, we execute the following two steps for each occurrence tagged therein:

6. Obtain stem information from external Glossary
7. Obtain Temporary Grammar Choices (TC_j) from stem information and ending information

The stem of a word is subject to many grammatical interpretations; the same is true about an individual detached ending. The intersection of the two sets of data, however, will usually be far smaller

than either set, but may still contain more than one grammatical interpretation of the word. We denote these possibilities as temporary choices for a given word. These now constitute the beginning of a string of facts which will replace the original text occurrence.

After all the text occurrences are processed in the manner described so far, we enter upon the next very important step:

8. Obtain Temporary Profile (TP) for each sentence by an iterative scheme

The purpose of the profile is to give the ranges of the clause and phrase formations within a given sentence.

We are then ready to attempt the following iterative technique for carrying out the syntactic integration of a sentence.

9. Make Foresight Predictions for future strings
10. Choose a TC_j as Selected Choice (SC) for the next string: Predictable and Expected; Unpredictable, Related Backward; Predictable but Unexpected
11. Record doubts about the SC in Hindsight 1 (H_1)
12. Store surplus TC_j in H_2 and H_3
13. Raise Chain Number, if SC is unexpected
14. Resolve, if possible, any previous Hindsights; indicate rearrangement of target order
15. Reduce Chain Number if resolution is obtained

The basic principle of our scheme rests in the series of predictions which each occurrence in turn makes anent future occurrences within a given clause or phrase. We predict from various sources:

(a) Grammar predictions are based on general principles which hold true for a large class of sentence elements; e. g., an adjective may be followed by another declinable word which agrees with it in number, gender, and case.

(b) Glossary predictions are based on the peculiar tendencies of individual words to govern certain subsequent occurrences; e.g., some verbs govern the dative case.

(c) Predictions from the physical appearance of the occurrence; e.g., a capital letter with period may be succeeded by a proper last name.

In relation to predictions, occurrences are either predictable or not. To the latter group belong the conjunctions, prepositions,

adverbs, particles, and punctuation marks. Every predictable occurrence is checked against the foresight pool, in which we keep all previous predictions that have not been fulfilled. The first temporary choice which fulfills such a prediction is selected as a link in our syntactic chain. If none of the temporary choices of an occurrence fits any of the previous predictions, we deem the chain to be broken. In that case we raise the chain number, which starts out as one, and we place an entry in one of our hindsight pools indicating this fact. It is expected that future occurrences will resolve this entry and in that case the chain number will be lowered. We also enter into our hindsight pools records anent doubtful choices and redundant choices. At the completion of an iteration which involves the entire sentence, we execute the next step.

16. Apply criteria for "goodness-of-translation"; repeat iteration, if necessary

The criteria consist of ascertaining that no unfulfilled predictions bearing high-order urgency numbers remain in the foresight pool, and that the final chain number equals 1. If either of these criteria fails, another iteration is attempted. When either a seemingly successful iteration is achieved, or the preassigned number of iterations have been executed without success, we execute the last two steps of the routine:

17. Rearrange target information as indicated
18. Print target sentence with signals
    a. Success or failure
    b_1. If success, number of hindsight entries
    b_2. If failure, type of difficulty encountered

Would interested readers be good enough to submit difficult Russian sentences to the above analysis, and indicate to us wherein the method fails?

As regards challenges of the second kind, we submit a list of the remaining six difficulties, which we have not yet been able to overcome.

(vi) Corresponding to an occurrence in the source sentence, more than one target word is listed in the glossary. Polysemy is, of course, recognized as a most formidable obstacle to faithful translation, whether human or mechanical. Hilarious (or heartbreaking, depending on your point of view) malapropisms can be cited by the

score to uphold the conviction of many linguists that the MT task is a hopeless one. Our faith in the inventiveness of the human brain makes us reject such gloomy forebodings.

(vii) The source sentence is grammatically incomplete. Such a situation is frequently the result of carrying on the thought from one or more previous sentences. To succeed, any MT scheme will have to be able to transcend the boundaries of a sentence (or a paragraph, or a section).

(viii) The source sentence contains ambiguous symbols. Since we are planning to confine our efforts to mathematical texts, such occurrences will be legion.

(ix) The syntactic integration of the source sentence results in an ambiguity. It is often of a type that could be resolved by semantic considerations; but, sometimes, it is inherent and thus not removable by any process.

(x) A combination of difficulties are listed in this category. They are quite annoying but fortunately rare: misprints; grammatical errors; localisms; peculiar nuances; comments based upon the sound (or the spelling) of source occurrences, such as puns whose sense it is impossible to render in the target language.

We should be extremely grateful to any of our colleagues who would take the trouble to outline his scheme for dealing with any of the above difficulties, and thus save us countless hours of unnecessary labor.