RUSLAN - AN MT SYSTEM BETWEEN CLOSELY RELATED LANGUAGES

Jan Hajič
Výzkumný ústav matematických strojů
Loretánské nám. 3
118 55 Praha 1, Czechoslovakia

ABSTRACT

A project of machine translation of Czech computer manuals into Russian is described, presenting first a description of the overall system structure and concentrating then mainly on input text preparation and a parsing algorithm based on bottom-up parser programmed in Colmerauer's Q-systems.

INTRODUCTION

In mid-1985, a project of machine translation of Czech computer manuals into Russian was started, thus constituting a second MT project of the group of mathematical linguistics at Charles University (for a full description of the first project, see (Kirschner, 1982) and (Kirschner, in press)).

Our goals are both practical (translation or re-translation of new or re-edited manuals for export purposes within the COMECON countries, of an estimated amount of 500 to 1000 pages a year) and theoretical (we wish to verify our approach to the analysis of Czech and to develop a theoretical background for translation between closely related languages such as Czech and Russian). The project is carried out by VÚMS, Prague (Research Institute for Computing Machinery) at the Department of Software in cooperation with the Department of Mathematical Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.

Input texts

The texts our system should translate are software manuals to VÚMS-developed DOS-4 operating system which is an advanced extension to the common DOS. The texts are currently maintained on tapes under the editing and formatting system PES (Programmed Editing System). This system allows for preparation, editing and binding-ready printout using national printer chain(s). Texts are stored on tapes using an internal format containing upper/lowercase letters, editing & formatting commands, version number/identification, info on last-changed pages etc.; most of this can be used to improve the overall translation quality. On the other hand, part of it is somewhat confusing and must be handled carefully.

By now, we have access to 65 manuals on tapes, containing about 12.000 pages (approx. 1.500.000 running words - 53.000 different word fomrs). The complete documentation covers 78 manuals and is still growing.

The overall structure

RUSLAN is a unidirectional system dealing with one pair of languages (SL - Czech, TL - Russian). We adopt a transfer-like translation scheme (in the sense we do not use any intermediate pilot language), but with many simplifications due to the close relationship between Czech and Russian, so that it belongs to the so-called direct method (in the sense of (Slocum, 1985)).

The translation process itself is to be carried out in batch (we have to respect the hardware available). This means that no human intervention is possible during the process. Nevertheless, our aim is to obtain high-quality results which would require usual post-editing only. No human pre-editing is contained in the system design.

The translation unit is constituted by a single sentence. Thus, the recognition of sentence boundaries is a part of the preprocessing.

For the time being, a treatment of ellipsis is not provided for, but a modification of the analysis is being prepared to account for cases (not very frequent in the translated manuals) where information necessary for an appropriate translation should be looked for in the previous sentence(s).

Translation steps

RUSLAN performs following steps to obtain the translation of a given (part of a) manual:

(1) The text is "punched" from a tape, to "visualize" all embedded editing & formatting commands;

(2) Fully automatic preprocessing follows, which includes:
- national & special characters conversion & coding
- sentence boundaries recognition

(3) The Czech morphological analysis (MA) is performed, followed by

(4) the syntactico-semantic analysis (SSA) with respect to Russian sentence structure, for each input sentence separately.

(5) The representation obtained in the previous step is converted into Russian surface word list in an appropriate order simultaneously performing some TL-dependent changes.

(6) Then, morphological synthesis of Russian (MSR) is performed and at the same time synthesized words are decoded and put out along with preserved editing & formatting commands, and at last

(7) the output is saved onto a tape under the PES system again.

The resulting text can be then easily printed and corrected using PES editing facilities.

Some more details

Since the overall structure of RUSLAN does not differ considerably from the existing MT-systems, we will concentrate ourselves in our paper on some interesting details.

ad (1): Getting a text out of the tape

This function is performed by means of PES "punch" command only. Internally coded words and commands are converted to card-like character format, so they can be read easily by other programs. This step is processed separatelly because we want to achieve the maximal hardware and operating system independence possible.

ad (2): Preprocessing

True words and punctuation are recognized and coded using alphanumeric characters only. Special characters (such as /, +, =, greek chars, etc.) and PES-commands are coded similarly, but they are handled as word attributes rather than as separate words.

The recognition of sentence boundaries proved to be the hardest problem of this stage. We have developed a special algorithm for sentence boundaries recognition, which takes editing commands and punctuation into consideration, as well as upper/lowercase letters in special positions. This algorithm is based on frames and features. Text is cut whenever the "End Of Sentence" condition is met. Such a condition is raised when one of the features of the next text element is found in the frame of the current text element.

Features assigned to each element are e. g. "beginning of sentence" - unconditional sentence boundary assigned to some PES commands, or "capitalized" - this one is assigned to the word starting with exactly one uppercase letter. Among other features we use there are "common word", "uppercase only", "number" and some other classifying PES commands.

Frames contain "beginning of sentence" in most cases; a more complicated situation arises when evaluating punctuation frames. Frames for ".", ";", "?" are created using quite complicated algorithms. Clearly, it is not possible to obtain 100% correctness without a deeper analysis, so we prefer (isolated) missing cuts to incomplete sentences. Tests showed only one missing cut every 100 pages of continuous text (introductory manuals), and every 30-50 pages in reference manuals; no incomplete sentences appeared anywhere in the sample. This looks promising, because missing cuts result in slowdown of analysis only.

ad (3): Morphological analysis

Since Czech is a highly inflectional language, this part is a little more complicated task than a MA for English. However, in the stage of MA of Czech we obtain much more useful information for the syntactico-semantic analysis.

MA is based on pattern unification. During the MA, the main dictionary is searched through to find all possible stems; ambiguities are treated in parallel during the next phase of processing.

ad (4): Syntactico-semantic analysis

SSA is the most important part of RUSLAN. Using Sgall's FGD as the theoretical starting point (for the most recent formulation, see (Sgall et al., 1986)), the dependency approach and data-driven parsing are the corner stones and valency frames are the tools of SSA. To control the combinatoric expansion, semantic features are used as additional constraints to the syntactic ones (for a more detailed account of

SSA, see (Oliva, in prep.)).

The result of SSA is affected by the TL-syntax - so there is no true separate transfer component in our system. In most cases, the need for changes can be resolved on the basis of the Czech sentence. A module is being prepared carrying out some minor restructuring (necessary e. g. for determining the word order and some instances of negation), which will be performed before the synthesis.

The close relationship between Czech and Russian helps us to leave many ambiguities unresolved and to allow the output to be as ambiguous as the input. We must resolve such ambiguities that would create multiple outputs in the TL, and select only one of them, but this is the case of only limited number of sentences.

ad (5): Generation

For the time being, no true TL-restructuring is being performed. During the dependency tree decomposition, morphological information is transferred from the governor to its dependent modifications according to agreement. The original word order is slightly changed when needed. An ordered list of words with morphological information and editing/formatting attributes restored is the output of this phase.

ad (6): Morphological synthesis

True words are processed by the MSR module to obtain their inflected forms. This module is capable of doing some word derivation (such as verbal adjectives). It is also responsible for orthographical changes (concerning prepositions and some pronouns) forced by the adjacent word(s).

After MSR, each word is decoded (including its attributes) to the PES-acceptable format and "punched" out. This is an inverse operation to step (2).

ad (7): Catalogization

Handled by PES solely, this is an inverse operation to step (1).

Implementation

All the testing is performed on the EC-1027 or IBM/370 systems at VÚMS (under DOS-4). The base of the system (steps 3, 4 and 5) is capable to run under the OS operating system as well.

Steps 1 and 7 are handled by special software, which is a part of the DOS-4 operating system. Steps 2 and 6 are written in standard Pascal (including the MSR module). Steps 3 to 5 are programmed in the well-known Q-systems, implemented through Fortran IV (G or H level). We use the Q-language compiler with the kind permission of its original author, prof. B. Thouin; some marginal changes were made in the Q-language interpreter due to the practical needs of our system. The only noticeable change is that complete graphs deleted formerly due to the CUL + DE + SAC mechanism are passed now (unchanged) to the next Q-system for further processing.

Maximal core requirement is estimated to 640KB (step 3 - dictionary), so it is possible to use even real-memory based systems. Secondary storage volume will be determined mainly by the dictionary size, since an average entry occupies 1000 bytes for the first operational version. We suppose that 10.000 entries will be sufficient for the first prototype. Dictionary search is performed using extended hashing scheme incorporated in the Q-language interpreter.

Elapsed time needed for translation depends on hardware and the time sharing coefficient. First test showed, that the widely-published speed of 1.5 mipw will not be exceeded. This converts to 3 sec CPU on our fastest EC-1027 computer, which will clearly suffice to translate up to the desired 50 pages a day.

Conclusion

In March 1987, steps 1, 2, 3 and 7 are fully developed and implemented, step 6 is implemented partially (morphological synthesis of Russian); it will be finished in mid-87. Steps 4 and 5 are under development. They have been separately tested since last summer, the manual on General Description of DOS-4 being the testing material. Translation of the first three pages is available now (performed by steps 3, 4 and 5). Simultaneously, dictionary entries (cca 7500 for the first, 87 version) are being prepared by external co-workers.

By the end of 1987, all steps (1) to (7) should be tested continuously at VÚMS. By the end of 88, RUSLAN should be able to translate existing manuals in quality worth postediting. When finished (1990), it should translate new software manuals in quality not requiring more postediting than human translations.

REFERENCES

Kirschner, Zdeněk. 1982. A Dependency - Based Analysis of English for the Purpose of Machine Translation. Explizite Beschreibung der Sprache und automatische Textverarbeitung IX, Charles University, Prague

Kirschner, Zdeněk. (in press). APAC3-2: An English-to-Czech Machine Translation System. Explizite Beschreibung der Sprache und automatische Textverarbeitung XIV, Charles University, Prague, 1987

Oliva, Karel. (in prep.). Programming a Parser for Czech - a Highly Inflectional Language, to be published in: Proceedings of the Conference on the Applications of AI, Prague, 1987

Sgall, Petr; et al. 1986. The Meaning of the Sentence in its Semantic and Pragmatic Aspects, Reidel/Amsterdam - Academia/Prague

Slocum, Jonathan. 1985. A Survey of Machine Translation: Its History, Current Status, and Future Prospects. Computational Linguistics 11: 1-17.