SOVIET PATENT BULLETIN PROCESSING: A PARTICULAR APPLICATION OF MACHINE TRANSLATION

Dale A. Bostad

ABSTRACT

Soviet patents are used for, among other things, gleaning information on developing trends in Soviet technology. Originally the patents had to be translated by human translators, indexed manually, and then entered into the CIRC II database for data manipulation and subsequent searching. This paper presents a description of the task, and of some of the technical problems that had to be overcome in changing to a machine translation system.

KEYWORDS: patents, Russia, Soviet, technology, machine translation, Russian, database, CIRC II.

This paper describes some of the processes involved in the data structure manipulation and machine translation (MT) of a specific text form namely, the Soviet patent bulletins *Otkrytiya, Izobreteniya, Promishlennye obraztsy,* and *Tovarnye znaki.* In the ensuing discussions, these bulletins will be referred to as OIPOB's.

BACKGROUND

Soviet patents provide a rich source of information on developing trends in Soviet technology and large numbers of them are translated and entered into a massive document retrieval data base called CIRC II. Soviet patents account for about 15% of this data base and as many as 40,000 units are selected for entry into it every year. CIRC II allows retrospective retrievals of documents based on subject content, facility names, equipment nomenclature, biographic data, geographic location, etc. The data base has detailed and complex formatting and indexing specifications for documents and requires a rather high quality of translation.

For years the usual method of preparing patents for entry into CIRC II was to have the patents translated manually by Russian translators, indexing for each patent was done manually, and then the patent with its appropriate indexing was keyed in on a terminal according to formatting specifications, and, finally, after a series of verification programs, the patents were loaded into the master data base. The entire procedure was time consuming and labor intensive.

THE TASK AND THE SOLUTION

The task to be solved was how to automate this laborious process of entering IOPOB's into CIRC II in order to achieve considerable gains in cost and timeliness. The proposed solution involved several sequential operations:

(1) Data capture of IOPOB's using the Kurzweil Cyrillic OCR or terminal input.
(2) Use of the Russian-English machine translation system for textual analysis and manipulation of data elements for indexing; machine translation of patent bodies.
(3) Development of an automatic indexing program for formatting data elements and establishing their interrelationships.
(4) Use of in-place data verification programs and final entry of patents into CIRC II.

From beginning to end the data was to be kept in digital form, passed from one stage to the next in tape media, with hardcopy printouts used only to verify or correct data. The overriding consideration was that the data had to rigorously conform to all CIRC II formatting and indexing specifications and that the machine translation quality had to be of such a level that it would satisfy accuracy and readability standards of a wide audience of users who accessed the data base.

In this discussion, I will only deal with the role of machine translation in the automated data flow process described above. And the central role of machine translation in the process should be emphasized because it is the keystone to the success of the project. The decision to go with machine translation was based on three commmonly quoted advantages offered by MT: best savings are made when large volumes of material are translated, and the patent project offered this; MT offers gains in translation turnaround time; and MT, across the board, offers savings in translation cost measure in dollars per thousand words. In addi-

Dale A. Bostad took his undergraduate degree in Slavic studies at the University of Zagreb, Yugoslavia in 1970, and the University of Poznan, Poland, in 1972. He has had seven years experience with the development of machine translation systems. He is presently a technical adviser to the Technical Translation Division of the U.S. Air Force.

tion to these reasons, we can say that the Russian MT system 1) for well over a decade had an established track record as a versatile, and efficient translation tool, 2) had linguistic algorithms that were sophisticated enough so that only minimal post-editing would be required, 3) had glossaries the scope of which could cover the terminology contained in the bewildering variety of Soviet patents in IOPOB's and, 4) could successfully and safely be modified for specialized text processing. The only question that remained was whether OIPOB's were in fact amenable to machine translation, textual analysis, and processing, and if so how this would be implemented.

MACHINE TRANSLATION'S ROLE

The task imposed on the Russian MT system was twofold: perform linguistic analysis and manipulation of certain data elements in patent headers; and develop logic to produce adequate translations of patent body texts. To accomplish this the general-purpose Russian MT system served as the baseline for the development of a special patent module PATSUBR, that would only be invoked when a PAT parameter was entered in the Job Control Language (JCL). Patent processing logic, then, was interwoven in the generalized logic of the system, with a special module written to handle those processes that had nothing to do with generalized machine translation.

PATENT HEADER TEXT MANIPULATION

Headers in Soviet IOPOB's contain different types of information. This information is sorted into elements and given a field designation number. For example, the patent number is the first data element in the patent header, designated (11), and is followed by a six-digit number. Other information fields follow, but the two most important ones are element (72) which lists the author or authors of the patent, and element (71) which lists the sponsoring facility or facilities involved in the development in a hierarchical ranking of subordination. Text processing of OIPOB header data consists of replicating all information in the given elements except element (72), where the names are transliterated into BGN (Board of Geographical Names' Standard Transliterations), and element (71) where special processing is necessary to prepare the data for the automatic indexing program.

Element (71) Analysis

Textual analysis of Element (71) had to be able to establish the hierarchical relations of facilities and determine parallel ranking of facilities where appropriate. Facilities with award data or honorifics designators had to be earmarked for index referencing; the city where a facility was located, if so indicated, had to be referenced; extraneous information in this element (71) had to be deleted; and certain types of textual massage had to be carried out. To do this element (71) was decomposed and synthetic elements designating subordination of institutions, the existence of honorifics, and the presence of cities were created.

Soviet patents provide a rich source of information on developing trends in Soviet technology and large numbers of them are translated and entered into a massive document retrieval database called CIRC II.

The first step involved descriptive analysis of a large corpus of patents (approximately 5,000) to determine recursive patterns and linguistic generalizations. The analysis revealed a wide variety of permutations in the presentation of facility information: patterns ranged from a single uncomplicated facility string to multiple parallel-ranking facilities, each with the potential of having higher ranking facilities such as,

> Ukrainskoe navchnogo-proizvodstvennoye obyedinenity tellyulozno-bumazhnoy promyshlennosti, Opytnoye konstruktorskO-teKnologicheskoye byuro Instituta metallofiziki AN Ukrainskoy SSR i Mariyskiy tsellyulozno-bumazhnyy kombinat.

As a generalization it was found that parallel-ranking facilities are in the nominative case and conjoined by commas or the coordinating conjunction *i* and that hierarchical facilities are related in ascending order of subordination from left to right, with capitalization and the genitive case the usual determining factors of a new level of subordination. Complexities in this general pattern were immediately apparent, for example in,

> Zavod-VTUZ pri moskovskom tridzdy ordena Lenina, ordena Oktyabrskoy Revolyutii i ordena Trudovogo Krasnogo Znamenii automobil'nom zavode im. I. A. Liyacheva

where a *pri* phrase marks a higher order facility. In this particular example, parsing matchup is aggravated by a long honorifics string separating the adjective, *Moskovskom,* and the noun it modifies, *zavode.* Aggravation of the parse by in-stream honorifics can also be seen in the example,

> Golovskiy filial Donetskogo ordena Trudovogo Krasnogo Znameni politekhnicheskogo instituta

where generalized logic would match the genitive singular adjective *Donetskogo* with the masculine singular noun *ordena.* Thus, it was apparent that honorifics had to be carefully demarcated, not allowed to enter the generalized parsing logic, and somehow be earmarked to refer to the correct facility they modified. One final consideration was noticed: when an honorific preceded a facility it took on the capitalization, thus removing the capital letter from the main adjective or noun of the facility.

Another problem arose with acronyms, since they may determine a discrete facility or they may belong to a previous facility string. Since acronyms, in most cases, are not declined, the general facility demarcation rules were not applicable. Finally, certain geographical adjectives in the genitive case also presented problems because they were exceptions to the general rule of facility determination: a capitalized noun or adjective in the genitive case.

The above description should give a good idea of the types of problems encountered in the analysis of element (71), although many other types of small linguistic problems occurred that go beyond the scope of this paper.

Element (71): Programming Solution

On the basis of the empirical observations of the patterns and exceptions described, linguistic programming was written to analyze element (71) and decompose it into synthetic elements. Reliance was based on existing parsing logic, but to account for the anomalies involved, special programming, including a superior facility table, negative semantic checks for geographical nouns, the transposition of capitalization from honorifics, etc. had to be implemented.

The logic demarcated the hierarchical facility strings. These were separated out, and given the designations (101), (102), (103) in ascending order of subordination. Parallel ranked facilities, if they existed, all received (101) designations. Element (106) was created whenever a geographic location (city) was specified. In addition to the creation of these new elements other text manipulations were performed:

(1) Honorific designations were deleted from the facility string, but their existence in the original element was flagged by creation of another field, element (105).

(2) Titles, initials, and first names were deleted in facilities named after people.

(3) Quotations were removed from appositional facility names.

(4) Downstream commas in long facility strings were removed, as well as the coordinating conjunction *i* and commas separating parallel-ranking facilities.

A typical example of element (71) in its original form and its breakdown into synthetic fields is:

commas separating parallel-ranking facilities.

(71) Krasnoyarskiy institut tsvetnyx metallov im. M.I. Kalinina, Noril'skiy gorno-metallurgicheskiy kombinat im. A. P. Zavenyagina i Institut Khimii Bashkirskogo filiala AN SSSR

(101) Krasnoyarskiy institut tsvetnyx metallov im. Kalinin

The central role of machine translation in the process should be emphasized because it is the keystone to the success of the project.

(101) Noril'skiy gorno-metallurgicheskiy kombinat im. Zavenyaging

(101) Institut Khimii

(102) Bashkirskiy Filial

(103) AN SSR

The next thing that had to be done was to put all elements created from (71) into BGN transliteration. And finally, any facility string in the locative or genitive case (the head noun and all adjectives to the left) had to be converted into the nominative case. This was done by putting relevant words through an adjective or noun paradigm table, where the endings were checked, the oblique endings deleted, and the correct nominative endings tacked on.

The success rate of textual analysis and breakdown of element (71) into synthetic fields has been about 99%. Because of the nature of the logic it is absolutely necessary that element (71) be input with 100% accuracy.

PATENT BODY ANALYSIS - ELEMENTS (54), (57)

Patent body processing and translation presented different problems that required different solutions. First of all, it should be noted that the syntax of OIPOB patent bodies is highly stylized and artificial, but that there are recursive patterns of phrasing in all patents that are easily identifiable. Most noticeable, however, is the fact that each paragraph consists of a single sentence and that the principle verb is a present participle and not a present active verb. Patents have been encountered with as many as 335 words in a single sentence. Such blockbuster sentences, especially in patents dealing with electronics and computer science, are so bizarre in their convoluted syntax that they simply defy rational description and, of course, they present severe translation problems to human translators, not to a mention machine translation system. It is necessary to intervene in these long sentences because (a) the maximum number of words per sentence that the system can handle is 145 and (b) the long complex sentences aggravate the readability and comprehensibility of the patents.

The decomposition of long sentences takes some patent-specific linguistic analysis and programming. I will briefly describe the procedure used. First, the entire sentence is read into storage and the first analytical program, EDIP, makes cuts for certain present and past participles that are immediately preceded by a comma in the sentence. The program looks for and makes cuts for ten specific participles (if present in the text), all of which must be in the nominative case, and all of which will fall within the range of a predetermined scan from the patent title (the exception being *otlichayushchiysya tem, chto* which can be located anywhere in the sentence and which is always the principle verb). Further cuts at this point are also made for six conjunctive adverbs.

Additional cuts are made later on after dictionary information has been read into the sentence and after basic parsing has been completed. For example, a sentence cut is made after the coordinating conjunction *a* if a verb or short form past participle is found in its clause. When the relative pronoun *kotoryy* is found in the genitive singular or plural *(kotorogo, kotoroy or kotoroyx)* a limited right scan is made for another *kotoryy* (also in the genitive case) that

must be preceded by at least one noun. If the right conditions are met, a sentence cut is made. Long strings of *kotoryy* clauses are frequently found in patents on electronics.

In all cases the preceding commas is converted into a period and then the participle or conjunctive adverb is capitalized to mark the beginning of the next sentence. In the case of *kotoryy* clauses the comma is deleted, the first word in the noun phase is capitalized and a period is placed before it. Finally, the coordinate conjunction *a* is set translated (given no translation in the text), the comma preceding it is changed to a period, and the first word after *a* is capitalized.

Because the knowledge is rather exact where the participles are placed in the sentence, it is also possible to go in and freeze the meaning of polysemous participles; if the participles occurred downstream, the system routines (multimeaning rules) determined the translations. Thus, for example, *imeyushchiy* was translated as include (usual translation is having) and *klyuchayushchiy* also took on the translation include (it means both switching on and including). The participles were then converted to finite reflexive verbs, 3rd person singular, present tense. The usual meaning of *kotoryy* (whose) was converted into noun phase + of the preceding. Other lexical and dictionary changes were made to correctly translate recursive phrases found throughout patent body texts.

OTHER PROGRAMMING CONSIDERATIONS

As stated, a special patent module, PATSUBR, was developed for patent processing. This module had six subroutines that did much of the specialized programming already mentioned. What was more difficult was entering the mainline logic of the Russian system to turn on switches for the specialized patent processing. This had to be done very carefully because of the complexity of the system's main linguistic logic. For example, certain procedures, including levels breaks, English synthesis, and rearrangement had to be no-opted (disallowed to function) for element (71). Special programming had to be carried out in the name/noun homograph routine to give precedence to the noun in all-capitalized titles of patents, and pronoun resolution had to be modified to give precedence in selection to inanimate translations of pronouns. All in all, a great deal of careful manipulation of the main system logic was involved for successful resolution of Soviet patent processing.

CONCLUSION

There are several MT systems dedicated to specialized text processing manuals or documentation. Our effort involved modifying a highly-developed general-purpose system, that had been used exclusively to translate full-text books and scientific articles, in order to do specialized processing and translation. The use of machine translation to carry out linguistic text analysis without translation is a unique feature, as is the programming to break down highly stylized prose into more palatable and readable units.

The results have proven to be a technological success; the machine translation of patents has produced translations of acceptable quality and the projected gains in productivity have been attained. ■

The decision to go with machine translation was based on three commonly quoted advantages offered by MT: best savings are made when large volumes of material are translated; gains in translation turnaround time; and, savings in translation cost measured in dollars per thousand words.

Funds to Improve Special Education

The University of Kentucky and the Fayette County (Ky.) public school system have received a federal grant to develop ways to use microcomputers to educate the handicapped children. The Department of Education grant is for $97,283.

During the two-year project, guidelines and methods for using computers in special education will be developed. Actual implementation is expected to start with the 1985 school year.

Multi-Lingual Input Provided for Text

Foreign language educators can now use Apple computers to teach foreign languages with The Professor's multilingual computer programs. Each program allows for the entry of text in Spanish, French, German, Danish/Norwegian, Finnish/Swedish, Hawaiian, Hungarian, Italian, Latin/Dutch, Polish, Portuguese, Turkish and Czech.

Multi-Lingual Language Teacher includes a library of drawings in color as an aid for teaching foreign languages. This program allows educators to create and operate their own programs for computer language instruction.

The Great Creator is a multilingual test generator program that produces multiple-choice, true/false, and fill-in-the-blank types of questions.

The publishers also offer several multilingual word games.

Disney's Educating Educators.

A series of three-day seminars created to help educators both in the classroom and in relating educational needs to civic groups, industry and government is being offered by Disney World in Orlando, FL. Twelve educator Seminars covering communications, marketing and advanced technology will take place in June, July and August, 1985, at the central Florida complex. For more information, write: Walt Disney World, Seminar Productions, P.O. Box 40, Lake Buena Vista, FL 32830; or call: 305/828-1500.