[From: *International Journal of Translation*, vol.13, no.1-2, Jan-Dec 2001]

Machine Translation Methods
Text Structure and Translator Work

LARISSA BELIAEVA
Hertzen State Pedagogical University, Russia

Critical aspect of society development in modern world is its scientific and cultural interaction, in which case both information technologies as a whole and machine translation become the most important facilities.

Nowadays, in the age of global communications, the need for a fast, accurate and cheap translation from one language into another has become very pronounced. This situation can turn critical when considering translation and interpreting in high risk technology domains. The discrepancy in Codes, Norms and Standards in various countries as well as delayed information exchange gives rise to increasing disagreement in high-technology and dangerous fields of common engineering interest. Furthermore, both fast and accurate translation means and translator opportunity to choose the proper means and to acquire special competence in their usage is the crucial point in implementing modern tools in translator work.

The main subject of this paper is correlation between the main problems and achievements of the present-day machine translation and translation memory systems and the opportunities to use such systems as a real tool for a translator. It means the necessity to account for the system pragmatics which demands to investigate different kinds of linguistic and extra-linguistic knowledge and take into consideration interference of the authors' mother tongues when they generate an English text in a special domain. One of the solutions is to give due consideration of the text structure and to use translation memory principles when a single text is processed in a machine translation system.

1. Introduction

During recent decades analysis of the situation in science and education has been carried out in the context of the information explosion to be in the nearest future. Under the conditions of strained expectation of this “act of nature” we in our country had overlooked the moment when we found ourselves in its epicenter.

Unfortunately, in spite of many years of preliminary discussions, the scientific and technical community in general appeared to be absolutely not ready to process the information flows in different languages neither technically, nor psychologically. The volume of current information, which knowledge is required to maintain basic, “mean-statistical” scientific level, is so huge and diverse, that even its preliminary analysis can take up all possible working hours.

At the same time it has long become clear, that only the use of computers when processing the text and document flows in various languages can allow the specialist to handle the information flow: to retrieve, attribute and use the necessary information.

Present situation when our country enters the common world and, more particularly, the European “global village” raises the question on analysis and consideration of information and documentation in diverse languages in a very special way. On the one hand, multinational community, now referred to as “United Europe”, carefully preserving all peculiarities of own languages and cultures, had chosen English language as lingua franca. Even preserving nine main European languages as official EC languages English language as it is remains the international communication language, from which and into which the main body of documents and literature is translated.

On the other hand, it should be kept in mind that modern international cooperation programs and joint scientific and engineering designs require not only common knowledge on the main avenues in scientific and technical development, but also coordination of international Programs. This situation can turn critical when considering the translation in high-risk technology domains. The discrepancy in Codes, Norms and Standards in various countries as well as delayed information exchange gives rise to increasing disagreement in high-technology and dangerous fields of common engineering interest. On the example of the seismic design, we can classify three groups of text materials that are very important from the international knowledge viewpoint and determine the safety level. They are state and international Codes, Norms and Standards; documentary flows on innovative techniques development. Volume of this documentation is so huge that it is not possible to organize its retrieval and translation with the aid of old conventional methods of human processing.

Under this conditions the need for a fast, accurate and cheap translation from and into English language has become very pronounced. With all these facts in mind, the problem of information flow organization and processing in different languages acquires a new practical significance. It is common place to advocate that this problem solving can be considered in language engineering aspect, the essence of which is development and/or adaptation of modern computer systems of natural languages text processing for specific research and technical problems, which are under investigation in various areas of science and engineering. Fast and correct translation of the proceedings of international conferences and symposia, current reports on the work of international research teams, coordination research meetings, harmonization of Norms and Standards in risky areas is a necessary tool for international cooperation and safety.

The peculiarity of the modern situation is related to the fact that a huge part of the international communication is in English language as lingua franca, which means that lexical and syntactic structures of the sentences used in these texts are frequently under strong influence of the mother tongues of their authors. At the same time all Natural Language Processing (NLP) systems are based on the assumption that source English sentence is grammatically and semantically correct. Interference of the authors' mother tongues, which is the urgent problem of language learning is now likewise urgent for NLP systems.

Advantages of practical NLP systems depend on our capability to solve this new bottleneck. Before proceeding to the possible solution of this problem let's consider the recent situation with NLP systems.

2. Natural Language Processing And Linguistic Automaton

Generalization of NLP system concept is an automated document processing system, named Linguistic Automaton (LA) which should function as a modular multilingual computer analogue of human verbal and mental processes and includes hardware, software, "lingware" and sometimes tutorware (Piotrowski, 1991;Piotrowski, 1994).

Such automaton is to be included in a system of information processing as a hierarchical system of NLP modules, each of which can function both independently and/or together with the others. These modules (subsystems) are designed to fulfill the following "intellectual" operations:

- Language Recognition, it means creation of a subsystem, the aim of which is attribution of any text from the data flow according to a particular language in which it is written. The list of possible source languages may be modified. The subsystem may be used independently or in connection with other LA subsystems. As a result of the operation of this module the input data flow is separated into files of texts in one and the same language. In case of a spoken language this module is to be used together with or after the next one:
- Language Recognition, it means creation of a subsystem, the
- Speech Recognition, which implies the tools for digitizing the speech into text form to be processed by other LA modules.
- Text Indexing, which means either processing a file of texts in the same language in order to separate it into files of texts, which belong to the same domain (scientific or user-defined) or to prescribe a special lexical index to the files from the data flow. The subsystem may be used independently or after Language Recognition Subsystem operation.

- Automatic Summarization, which means extracting the most important sentences and compressing the text up to a text summary, which consists of the most important sentences. The degree of text compression up to the summary may be determined in dialogue way on-line. The subsystem can be used independently or after any of the above-stated subsystem operation.
- Machine Translation, which means generation of a high-quality or rough (informational) computer-translated text. This function can be implemented in kinds types of systems. The first way is to create a translator workstation as a multilingual system to provide for high-speed and accurate translation of texts from one language into another in specific fields (or domains) of science and technology. The second way is creation of a computer tool fro the specialists, who don't know foreign language, this tool helps specialist quickly and with small expenditures to receive approximate (rough) text translation in the area of his interest, the translation, which is sufficient for understanding the information of the foreign language text.
- Dictionary Acquisition, which means automated support of lexicons (resident and automatic dictionaries) for different domains (Wilks et al., 1996).
- Computer Support of native (foreign) Language Learning, which will include all possibilities of the above-mentioned LA modulus.

The problems of creating Linguistic Automata can be considered as a long-term program of language engineering, the result of which shall be constructing a linguistic processor being, in fact, an Artificial Intelligence system. Modularity and hierarchical structure of such LAs determines the possibility of their creation by way of successive iterations even within the limits of each separate module.

In the context of the problem set above a consideration of the real state in machine translation development is of specific interest.

3. Conventional Machine Translation Approach

The objective of machine translation (MT) is to translate documents more quickly and cheaper than a human translator could do. As we consider a MT system as a practical one, which is designed, first of all, to provide convenient work for both a translator and a specialist in some technical domain, who doesn't know foreign text language or knows it not well enough, this aspect shall determine functionality of all solutions accepted. In Machine Translation Laboratory of Herzen State Pedagogical University of Russia we had been working on MT systems for more than 20 years. The first practical system SILOD had been created for English-Russian, Chinese-Russian, French-Russian, Spanish-Russian and Russian-English language pairs and had been realized on a mainframe. New PC-oriented versions (MULTIS, PC-SILOD and the last version for Windows - PRAGMATICS) have the same theoretical basis. The optimum MT organization requires a modular design that consists in realization of MT system as a set of non-rigidly linked modules. This modularity makes it possible to arrange an MT system as modules are ready and eliminates the data duplication. Besides, it allows the step-by-step solving of MT problems. Translation is regarded here as a multilevel process, where each procedure translates a component of the special level. It means that the source structures of each level are to be recognized, described and transformed into output structures, which may be modified on the next level in accordance with its structural features.

Thus the translation process is simulated in the system in question as a composition of lexical and semantic-syntactic translation processes. To create a practical system, we must orient its structure on the system pragmatics, which makes it necessary to investigate into different kinds of linguistic and extra-linguistic knowledge.

In accordance with all this each level of processing has its own functional value. Thus, when creating a practical MT system in our team (Piotrowski, 1991) we have accepted the following hierarchical levels of system implementation:

- automatic pre-editing of the source text;
- lexical and morphological analysis of the source text;
- contextual analysis and analysis of groups;
- functional segment analysis;
- sentence analysis;

- synthesis of target text;
- automatic post-editing.

Hierarchy of these levels assumes the possibility to realize the system from the highest levels down to the lowest ones, but not necessarily in succession, that is, with a net result received on each of them, that in its turn is related to setting the purposes and peculiarities of procedures organization on each of the levels. Thus a translation process is considered as a multilevel process in which every procedure accomplishes the translation of the components on each separate level.

Each procedure answers the purpose of building the components on every level, and connection between the levels is realized by transferring the parameters that represent the attributes of the components in a given bilingual situation. On the lexical level the source chain representing the source text is transformed into a chain of coupled source-target lexical units and is supplied with a set of attributes provided by the dictionary for that coupling. When the MT system uses a stem dictionary, the meanings of the attributes are defined more exactly and completely by means of the morphological analysis. All other levels of analysis are realized not on the source text as a whole but on the sentence-by-sentence basis. Thus on the group level the chains of lexical collocations are transformed into the chains of collocations of the source group-target group type, and so forth. It means that the source structures of each level are transformed into the output structures that can be modified on the next level in accordance with the structural features of this level. The translation process is simulated as a composition of the lexical and semantic-syntactic translation processes realized on the sentence-by-sentence basis.

From the possible strategies of MT the most widely used and efficient is the strategy of translation with transfer. The transfer approach is usually realized by means of lexical functional grammar. The main feature of the transfer approach is structuring the translation process into subprocesses and finding out the constituents of any subprocess. When the operational procedure of a special level is accomplished the system transfers the results to the next level.

Using the transfer method makes it possible to accomplish the switching from the subtrees which describe the source sentence structure on various analysis levels to the deep role structure, which is to be used as the basic grammatical representation, and then to the source subtrees. Under this condition using the transfer procedure just for local subtrees permits to receive a MT without simulation of complete sentence understanding, which is absolutely not necessary during MT by virtue of the fact that the deep semantic and syntactic interpretations can be left to the end user, i.e. to the specialist, who works with machine translation results.

The levels of practical MT system realization proposed here determine the peculiarities of its soft- and lingware. The last one is realized as a closely correlated system of automatic dictionaries (AD) and grammatical procedures (Beliaeva, 1992; Beliaeva, Bychkov, 1996).

The optimized MT system organization requires a modular design that consists implementing an MT system as a set of non-rigidly linked modules. This modularity makes it possible to arrange a MT system as modules are ready and eliminates the data duplication. Besides, it allows the step-by-step solving of MT problems.

One of the key moments in the MT problem solving is a noun group translation, in case of technical and scientific texts this procedure often means the terminology translation. The correct and accurate terminology translation is crucial for the quality of a machine translation result. Thus the level of context and group analysis acquires a particular importance. On this level the next procedures are to be implemented:

- determination of the noun and verbal group boundaries;
- context analysis of homonyms and solution of homonymy on the group level;
- transfer of groups and transformation of the chains of lexical pairs into the chains of pairs of source group/ target group type;
- modification of information on the collocations and separate words translation.

Procedures of this level are usually implemented with the help of the Augmented Transition Networks method. A lot of complicated problems are to be solved here, in fact, the main goal is to find a predicate or a series of predicates. Then on the next level it would be possible to determine position of a subject or subjects. Thus, the homonymy of Past Indefinite and Participle 2, verb and noun, Gerund and Participle 1, etc. are to be solved. At the same time, the boundaries and the structures of the noun groups are to be established and transferred just here (difficulties in trying to solve the problems of homonymy of 'flying planes' type are well known).

It must be specially emphasized that any syntactic and semantic analysis level should have a special logical procedure, the goal of which is the selection of one special type of case, part of speech, function of a group or sentence structure. It means that after the analysis of the chains of the source words codes and the hosts of noun and verb groups, the program is to solve the remained homonymy of verb forms and to fix a predicate, if no predicate was found on the basis of context analysis. Now this logical procedure is based on the hierarchy of candidates, which is to be established on vast investigations of possible combinations of the codes. Implementation of a logical solution procedure is a method to reduce the quantity of possible parsing results especially on the background of the interference of the author's mother tongue (it should be noted for example that absence of the plural number characteristics in Japanese generates permanent mistakes in grammatical agreement in number in the English texts written by Japanese authors).

At the same time on this level we are faced with a complicated terminological problem: with the problem of determination of the noun phrase (NP) structure, especially for the analytical languages. This is not only the common bottleneck for a MT system and language learning, but the problem which is immediately related to the interference of the authors' mother tongues. Investigation of texts written in English by authors of different origins (from Japan, Korea, France, Belgium, Germany, Argentina, Chile etc) shows that NP-structure in the texts of different authors correlates with the structure of the noun phrases or the terminological items or collocations with the same meaning in the author's mother tongue. Thus the problem of NP analysis and its ambiguity solving is complicated by this permanent violation of English grammar.

At the same time, in English language absence of morphological indications of gender and case for NP elements which could show grammatical agreement and, correspondingly, dependencies structure of a NP makes it impossible to establish the "host" for the adjective or a set of adjectives in the preposition to a chain of nouns (see, for example, a noun phrase 'constant amplitude deformation cycle'). In such cases a semantic approach (modeling extralinguistic knowledge of the domain, in which MT system is realized, with the help of thesaurus or semantic nets) can be implemented. This approach could be based on vast investigations of possible relations between both the main concepts of the domain and the items of the linguistic database. Development of such a thesaurus or a real semantic net is not only extremely laborious but is rather time-consuming. But the most serious disadvantage of this approach is that an unambiguous solution in many cases can't be achieved. For our example, in a semantic network there would be shown relations between all the nodes 'constant' and 'amplitude', 'constant' and 'deformation', 'constant' and 'cycle' and it is impossible to use this information to establish the dependencies structure of the NP.

The quality level of the machine translation results received with the help of a MT system depends on its adjustment to the problems of a certain user. But even with the insufficient level of such adjustment a MT system application permits to lower the expenditures on translation up to 20% from its initial cost, let alone speed of its performance even if a translator works as an editor of the received machine results (Krauwer, 1996).

Recently we can see a new tendency, which relates to the use of MT systems in large translation firms and centers - with the advent of Translation Memory systems and with a new interest to the quantitative analysis in the domain of MT.

4. AUTOMATIC DICTIONARY ITEMS IN LINGUISTIC DATABASES

Dependence of the database structure on the knowledge domain and main task of a MT system and any NLP system and as a consequence, the necessity of AD adjusting to the domain peculiarities is now mutually recognized. The same refers to the volume of a NLP-system database. It is now absolutely clear that creation of a practically usable expert system makes it necessary to design a huge database, items of which can represent the main concepts and conventional terminology of the domain in question. Not less than 95% of the source text items are to be distinguished and described with the help of a database if the expert or NLP system is oriented for a practical use.

Naturally, particular volume of a database depends on the typology of the source language and the chosen procedure of morphological analysis, the aim of which is high-speed and accurate identification of the source text word-forms with the help of AD.

Choosing a particular form of AD items depends on two main parameters:

> time-consuming parameter, which shows the time necessary for a text word-form identification;
>
> stability of morphological procedures which are to ensure word-forms identification and prevent improper identification.

Thus, we know that the first task of AD in any NLP or expert system is text word-form identification, procedures of which depend on the chosen type of machine morphology and as a consequence on the type of AD items. The choice of AD item is determined both by word- and form-building principles different in specific languages as well as by the representation of semantic items of a text. Besides, the choice of a basic dictionary item is determined by the tasks of NLP system.

Our experience in designing AD for typologically different languages has shown that for analytical languages a dictionary created as a set of separate word forms is most expedient as it makes it possible to increase the speed of the system while the growth of the dictionary volume is negligible. For synthetic languages adoption of special computer methods of morphological analysis is equally expedient. In this case the machine stems are considered as the heads of dictionary items (DI) and AD is to be provided with machine morphology. The same approach is needed for speech recognition database as the stem (the initial part of a word) can be recognized more correctly than the ending.

In order to reduce the memory volume for AD location it is advantageous to use an artificial morphology transformation, i.e. the agglutinative morphology. The essence of the latter consists in the process of the separation in any word usage a machine stem and an affix "sticking" to it.

The concept of the machine affix separation made it possible to elaborate principles of machine morphology creation, universal for many inflective and agglutinative languages (for example, now these principles are justified for all Roman languages, Russian, German and Greek languages and for some part of Finnish language morphology). Machine morphology is formed as a set of paradigms - machine affix chains. Each typical paradigm correlates with the grammatical characteristics of the stems and the word formation mode. The link between a machine stem and a paradigm is realized with the help of a special code, which characterizes all possible word forms which can be generated from the stem in question.

The use of this machine morphology allows to realize any word-form recognition and generation procedures in accordance with the lexical and grammatical characteristics formed in the course of NLP, and to make such procedures universal. For example, the Russian stem dictionaries elaborated for machine translation (MT) purposes permit to identify automatically the text words with dictionary items and to ascribe their morphological characteristics within the accuracy of case homonymy.

The result of morphological analysis, which is received with the help of AD and special lexical and morphological analysis algorithms, is a source for NLP algorithms. Irrespective of the use of machine morphology any AD includes both word-forms and stems, because always there are some cases that make use of word-forms advantageous (for example, it is expedient to include all forms of modal and auxiliary verbs in AD of any language). In case of speech recognition these elements of text are usually not accented, thus they are to be recognized as a not-segmented integrity.

Requirement to include into AD both separate lexical units as word-forms and stems and combinations of such elements (machine phrases) is recognized now by all database designers. But the bottleneck of such automatic machine phrases dictionaries (AMPD) lies in the necessity to establish for any database the following: typology of machine phrases; method of their recognition in course of text analysis; method of AMPD storing.

The problem of AMDP is connected with the fact that new and important notions, in all contemporary languages, are often expressed by means of phrases. Therefore, when creating a linguistic database it is essential to elaborate a frequency dictionary (FD) of phrases for every application domain alongside with the traditional FD of words. Such domain-oriented FD of phrases can be used to compile huge AD of phrases and words and to solve the problems that were listed above. In case of speech recognition system it is necessary to take into consideration the accent structure of a machine phrase - is it a large phonetic word with one accented syllable and reduced ending, which can include several words, or a machine phrase is pronounced as a set of words with special prosodic features.

From typological point of view it is possible to determine the following kinds of machine phrases (MP):

- icon MP, i.e. unchangeable linear combinations of word-forms, functional, semantic and syntactic characteristics of which do not depend on the context (in all languages we can find composite prepositions, conjunctions, adjuncts etc. as such MP) as usual they are large phonetic words;
- icon changeable MP, which are represented by linearly continuous sequences of words, functional characteristics of which depend on the syntactic function of a MP in the sentence (see terminology in any domain);
- conventional iconic MP, which are represented by unchangeable sequences of words, functional and semantic characteristics of

which depend on the context, in particular on presence of a punctuation mark, which can isolate a parenthetic clause; discontinuous MP, which are represented as sequences of words, be used verbal phrases. As to the morphology the discontinuous MP can be both changeable and unchangeable [4].

5. Dictionary Item Structure

Thus, any linguistic database, which can be a part of NLP system or a special entry to knowledge or terminology database of any expert system, at least includes source word dictionaries, which are organized both as dictionaries of word usages and dictionaries of stems, source phrase dictionaries and machine morphology for different languages,

AD is the key part of any MT system, because its linguistic algorithms and software could be realized just on the basis of the information stored in the AD. So a special consideration must be given to the volume of information to be ascribed to any AD element and to the mode of its storing in the AD and extracting from it.

Experience of practical MT system designing had shown that it is impossible to elaborate a complete structure of DI for any MT system ad hoc, at once and for all theoretically possible situations. Even if the procedure of creating a word portrait is attractive for a linguist, in reality we must include in the DI only the information justified by the algorithms realized.

Naturally, such approach must be added with creation of special procedures and conditions that allow to complement any DI with new information which is acquired as necessary.

In our PRAGMATICS version of MT system which is designed as a translator workstation system with functions of MT, language identification, synonym and thesaurus information access, residential dictionary implementation etc. any AD that characterizes a specific language has a universal structure of dictionary item and special machine morphology. All the source language ADs have the same function and a united scheme organizations.

This scheme allows to unify such procedures of the source language text processing as a selection of minimum text units, the morphological analysis, the identification of the text with AD items, the organization of the dictionary information file. Any lexical unit (LU) in AD acquires a description on the morphological, syntactic, semantic and functional levels as an appropriate characteristic set. The basic version of the system includes dictionary items (DI), which are defined as a set of the following characteristics:

head LU as it is: a stem, a word-form or a MP (this part can be added with phonetic record);

lexical and syntactic code (LSC), which depends on the typological features of the source language, its grammar and parsing or semantic analysis algorithms which are realized in the system in question;

translation, which can be stored as system of references to the corresponding target language items (stems and lexical and grammatical characteristics).

In our system the Russian language is used as a metalanguage for source text definition as well as the target language for many MT systems. The unity of the target language enables to unify its definition for all NLP systems from foreign languages into Russian and to unify the procedures of morphological synthesis of Russian word-forms.

When we design a MT system for Russian texts processing the morphological analysis procedures of are unified as well. In any case machine morphology definition of the Russian language constitutes a separate module and can be used in all versions of NLP systems.

6. Linguistic Databases Software and Maintenance

When designing linguistic databases for practical NLP systems we differ two types of databases:

linguistic database (LDB), which include ADs in a convenient

form and facilities of its updating and modification, which are oriented on the problems solved by system designers (linguists, knowledge engineers etc.) and
special dictionary files, which are results of special program-simulated conversion of the linguistic database into a format intended for the system software.

As a result of such approach the format of the dictionary files can be changed as the system is developed, but the linguistic database (when the service utilities are advanced) can progress independently without obligatory rearrangements.

When designing a LDB for translator workstation we take into account a linguist or a knowledge engineer tasks and to pay special attention to convenience of their work and easiness of LDB updating.

Requirements to the production rate and processing speed are not critical, the last must only correspond to the rate of operator work on a computer. But the convenience of the operator work is more than important. Thus the structure of such a LDB is oriented on the user convenience requirements. An LDB is supported by a convenient system of helps that ensure dialog mode of any field filling. These helps are designed in such a way that a user without special training can create a new DI (even not knowing peculiarities of NLP system). Besides, LDB system utilities support the possibility to look-up, update or modify information in any dictionary item field, as well, to look-up the LDB on the basis of any characteristic in any field. Several LDBs can be merged, compared, their common part can be extracted, etc. On the base of these service functions there can be created not only a set of domain-oriented LDBs but a universal LDB as well.

In this database special service functions that make it possible to create in a semi-automatic mode a new domain-oriented LDB on the base of text sample processing and extraction of DI from a universal LDB or from several domain-oriented LDBs are realized. Huge collection of service functions slower the processing rate of such a LDB but this rate is enough as to ensure convenience of real-time operation.

When the DB is included in the NLP system the requirements to the production rate and processing speed increase critically. A major part of time such DB spends on data look-up and reading. Addition of lexical information or correction of dictionary items of LDB either is not performed or may be performed only if necessary as a special session.

Under these conditions it is not expedient to use as such database the LDB which has special functions of information accumulation and modification and is to be convenient for a linguist and not the NLP system. Besides, it should be emphasized that the problem of disk space volume to be allocated for LDB is of great importance even nowadays when computers with huge hard disks are available.

Thus, under the approach in question it is absolutely necessary to create a special format of LDB that meets all the requirements established above.

5. Quantitative Analysis Methods and Machine Translation

Machine translation is one of the domains where the quantitative methods can be used both at the stage of practical MT system design and at the stage of investigations of translation processes and peculiarities.

When creating a practical MT system the quantitative analysis shall be implemented to solve the following problems:

determination of both automatic dictionary (AD) structure and dictionary items structure on the basis of a statistical study of word distribution in the domain-oriented texts;
selection of basic terminology to be included into the AD on the basis of different distributions in a sample of texts;
investigation of syntactic models of sentence structure for restriction of the parsing methods to be implemented in the syntactic component of a MT system;
statistical approach to text structure study for determination of the text peculiarities to be allowed for in a MT process.

Recently a special kind of NLP systems is actively used in translation practice. They are the systems with a Translation Memory (TM) (Merkel, 1996). The main idea of such TM systems is storing in a computer memory the aligned corpora of the source and target texts. These corpora could be formed by a free-lance translator or by a translator team during their translation work. Accumulation of the aligned texts makes it possible to reproduce the translations that had been done earlier. Such TM systems can be treated as "learning" systems since when using them any sentence is to be translated only once. During a new text translation the procedure is as follows: the TM system tries to find in the corpora the same source sentence. If such sentence is found the source sentence acquire its translation, otherwise the system "tries" to find a sentence, in which most parts coincide with the sentence to be translated. Degree of such coincidence and procedures of its evaluation depend on the TM system used.

The statistical translation technique is based on the stochastic language model and uses little linguistic and no semantic information. The main.principle is the probabilistic analysis of the aligned corpus which ensures a good quality translation of short sentences (Brown et al.,1993). Naturally such systems are very efficient in the translation centers of large firms where the descriptions of the products under production change only in small details.

Thus the main problems to be solved when creating a TM systems are related to the statistical assessment methods and the alignment procedures. When solving MT problems with statistical methods no sophisticated parsing algorithms are to be used. Some robust parsing methods can be applied only during pre-alignment process. Under such approach any sentence is considered as a chain of items, each of which can be translated into the target language on the item-per-item basis. The problem is to find boundaries of these items - sentences, flexible or rigid collocations or even words and to establish their possible translations on the basis of bilingual corpus.

Thus for a TM system the bilingual corpus must be aligned sentence-by-sentence. This pre-aligned corpus is used as a reference database. During the work with a TM system the translation of any new source sentence (whether done by the system or a translator) is added to this database. Besides the alignment can be done on the basis of words and word groups. In the aspect of this paper the most interesting is the word group alignment which is important for identification and translation of terminology. Word group alignment is based on the assumption on the collocation semantics that each collocation is unambiguous in the source language and has a unique translation in the target language (Smadja et al., 1996). In spite of the fact that this assumption is too strong even for a bilingual corpus for one and the same domain it can be very useful if we conjugate the MT and TM systems. The last kind of the systems are now quite popular but the peculiarity of the approach proposed is to implement in a MT system creation the principle of TM not for a bilingual corpora but for a separate text translation.

6. PRAGMATICS-WORD+1.5 MACHINE TRANSLATION SYSTEM AS A TRANSLATOR WORKSTATION

New version of our Word-version of the SILOD MT system is PRAGMATICS, which can work as independent system or as a built-in library Word+1.5. The system is realizes as translator workstation and can be used for text translation and editing, for language learning and translator training. This system has been under development in Machine Translation Laboratory of "Hertzen" State Pedagogical University of Russia (St. Petersburg)

SILOD-Word+1.5 Machine Translation System is a system which maintain linguistic support of translator work with the Microsoft Word editor (versions 6.0,7.0 and 8.0.). The system user gains access to the following functions:

> Machine translation. SILOD-Word+1.5 includes Machine translation system for translating from English into Russian and from Russian into English using base dictionaries for various domains (business, digital communication, seismicity, nuclear power plants, medicine etc). Translation can be carried out in batch or dialogue mode. Under the dialogue mode a user can receive per-block or per-paragraph translation. Both the source

and the target texts may be looked up and, if necessary, they can be easily edited with the help of a convenient built-in editor, which permits to work both with separate words and blocks. The system has the possibility to create special user dictionaries and to set the queue for the system reference to the dictionaries when searching the proper word. The process of dictionary creation itself is simple enough and comprises a dialogue between a user and the system, during which the user is proposed to answer a series of simple questions, as a rule this answering does not require any special knowledge.

Spell-checking. This function makes it possible to perform spell-checking for the text in 2 languages.

Finding synonyms and word-forms (for the Russian language). This function ensures the possibility to find a synonym in case the translator is not satisfied with the translation or the source text style. Furthermore, this function helps those who are not fluent in Russian to find the stem or to learn the word-forms for the word in question.

Using resident dictionaries and thesauri makes it possible to extend the limits of AD and find new words or new translation options for a word or a collocation. If this function is used for a word not included in any of the system AD it is possible to create a new dictionary item in dialogue and save it in proper basic or user dictionary.

These built-in functions of the system in question ensure convenient work of translators. But, nevertheless, any translator should choose the proper system on the basis of his aims and text peculiarities. To show the MT scope of the system I'd like to present the results of an experiment which is not absolutely correct from the linguistic point of view. We took a part of real text in English, translated it into Russian and then translated the MT result in English (all three specimens can be seen in Appendix).

7. Text Structure and Machine Translation System Competence

It is necessary to remind that the real unit of MT process is not a text but only a sentence. We all understand that it is a simplification, since the proper text translation can be done only on the basis of its processing as a whole. But in the case of MT, the sentence is the main unit in any operational system. This restriction is a very severe one from the terminological point of view. Thus the main bottleneck of different MT systems is the sentence-by-sentence analysis of any text. It means than when translating a text the MT system begins to translate the next sentence as if it is the only sentence to be translated.

Investigations of mental activity during both text perception and generation show that a text itself is a composite reflection of the author's knowledge and intentions in which linguistic competence is realized on the basis of a sequence of the predication acts. The hierarchy of the predication acts or the chunks of meaning recognition is determined by the author’s competence: professional domain knowledge, language competence and personal intentions. As a consequence, the text is characterized by such features as cohesion, integrity, completion, peculiarities of its component structure, etc. Furthermore, a text reflects the author’s notion of a situation and its elements, their relations, dependencies and hierarchy.

A special field for investigation of text structure and realization is a comparative study of the texts that are characterized by comparable volume, restricted formal structure, the same domain and language.

The restriction of a domain, volume and formal structure (for example, rigid sequence of title, abstract, sections and subsections) can eliminate the influence of different extra-linguistic factors: different types of terminology, accepted in various domains, dimensions of texts, orientation on a certain type of recipients, etc.

A comparative study of a set of such texts (a sample from international conference proceedings) has shown the real distribution of terminological elements, quantitative and structural peculiarities of technical texts of a comparable volume as integral units (instead of a text sample), regularity of noun phrase structures, etc. Such results are important for creation of real binary dictionaries for scientific-technical translation and for recommendation on the translation methods.

Investigation into the lexical and NP structures of international conference proceedings shows that average length of such texts is 1699 words, average number of different word is 490, average number of collocations with a noun as a kernel (noun collocations (NC) which constitute the basis of a NP) is 236 and average number of different noun collocations is 129. That means that such texts are "packed" with NCs and repeating of one and the same collocation is quite high. Analysis of the sequence of the NCs in such texts shows that in the beginning of the text we can find the most frequent NC with the simplest and proper structure, the boundaries of which are accurately shown by the authors by means of articles, prepositions, etc. As the text approaches its end all these features can be lost, the NCs are united in more long and sometimes grammatically-incorrect NCs and NPs, recognition of which would be a problem for any MT system (sometimes they are a real problem even for a translator).

This analysis gives the possibility to suggest a "psycholinguistic" basis for conjugation of MT system and TM principles: in the process of MT the information, which can be received on the basis of the whole text analysis should be used. This approach is based on the formal indications of the author's intentions that are reflected both in the text structure and in the composition of different NP with the same constituents. Investigations of text structure in terms of NP composition in different domains (medicine, seismic isolation, space facilities, power plants construction) show that dependencies structure in NP with 3 or more constituents can be obtained from the nearest context: 2-component NPs would show the accurate relations relevant for this special text.

8. Conclusion

Comparative study of the texts written in the English as a lingua franca can clarify a question of mutual influence of lingua franca and national languages of the authors, which are forced to create a text with preset parameters of subject, structure and volume in a foreign language.

Statistical study of the results of MT is a way of text analysis in which synergetics of mental and linguistic processes can be revealed most explicitly. Thus such comparative statistical analysis of MT system and human errors can highlight peculiarities of translation process as it is.

REFERENCES

Beliaeva L. 1992. Linguistic Data Base for Natural Language Processing. In: Conf. Terminol., Standard. Technol. Transfer. Proc. 1991-07-02/06. Beijing (China): Science Press, pp. 387-392.

— 1995. Machine translation and translation work. In: Translation and Meaning. Part 3, Maastricht, pp.241-247.

Beliaeva L. and Bychkov V. 1996. Automatic Dictionaries as Models of Professional Competence in Linguistic Automata and Speech Recognition Systems. In: Proc. Intern. Workshop: Speech and Computer, St.Petersburg, Russia, pp.45-50

Beliaeva L., Piotrowski R. and Sokolova T. 1990. Principles of Linguistic Automaton and their Information Bases Design. In: TKE'90. Second Intern. Congr. on Term. Knowledge Engin. 2-4 October 1990, University of Trier (FRG). Assoc. Terminology and Knowledge Transfer. Intern. Inform. Center for Terminology (Infoterm),pp.419-425.

Brown P.F., Della Pietra S.A., Della Pietra V.J., Mercer R.L. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. In: Computational Linguistics, vol. 19, No2, pp.263-311.

Expanding MT Horizons. 1996. Proc. of the Second Conf. Assoc. Machine Transl. in the Americas. 2-5 October, 1996, Montreal, Quebec, Canada. 289pp.

Krauwer St. 1996. On the Role of Machine Translation in the Multilingual Information Society. In: Proc. Intern. Workshop: Speech and Computer, St.Petersburg, Russia, pp.45-50

Merkel M. Checking Translations for Inconsistency - A Tool for the Editor. In: Proc. of the Sec.Conf. of the Assoc. for MT in the Americas. Montreal, Quebec, Canada, 1996, pp. 157-167.

Piotrowski R. 1991. Development of a Linguistic Automata in Speech Statistics Group. In: Autom. Docum. Mathem. Linguist., Allerton Press (USA), vol.25,No 9, pp.12-18.

— 1994. Psycholinguistic basis of the linguistic automaton. In: Intern. J. Psycholinguistics (Osaka, Japan), vol. 10, pp. 15-32.

Smadja F., McKeown K.R., Hatzivassiloglou V. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. In: Computational Linguistics, vol.22, No 1, pp. 1 -38.

Wilks Y.A., Slator B.M., Guthrie L.M. 1996. Electric Words: Dictionaries, Computers, and Meanings. Cambridge, MA: The MIT Press, 1996. pp.288.

APPENDIX

Examples of English-Russian and Russian English MT translation of real scientific text.

English Original Text

Current Status of Seismic Isolation Technology in the United States

James M. Kelly
Professor of Civil Engineering
University of California at Berkeley
Earthquake Engineering Research Center
Richmond, California 94804

Abstract

Seismic isolation is at the present time in a very active state of development. Many new types of isolation systems are being explored and elastomeric isolators, the system which has been employed on almost all isolation systems completed to date, continue to undergo improvement. At least one dozen large projects, either new or the retrofit of existing buildings, have been completed and design studies are underway for at least another one dozen large projects.

A large experimental research project for isolators with nuclear reactor application has been carried out over the past few years at EERC. This program has involved shake table testing and the testing of full-scale and model isolators. A wide variety of isolators have been tested including low-shape factor, moderate-shape factor, and very high-shape factor elastomer bearings.

The range of elastomers that have been tested include low-damping, high-damping, and very low-modulus compounds. Full-size and model isolators have been tested to failure in several failure modes and the safety margins for isolation systems have been established. The test results have shown that properly designed and manufactured isolators for nuclear reactor applications can sustain levels of loading beyond any possible seismic input and demonstrate that failure of an isolation system cannot occur before failure of the isolated structure. Thus, the use of isolation can only have beneficial contributions to the protection of nuclear facilities, internal piping, and equipment.

The presentation will review the latest development in the implementation of base isolation and describe the results of the test program for its application to nuclear facilities.

Russian Machine Translation

Russian Machine Translation

Нынешнее состояние Технологии Сейсмоизоляции
В Соединенных Штатах Америки

Джеймс M. Kelly
Профессор Строительной части
Университет Штата Калифорния в [при] Berkeley
Исследовательский центр сейсмостойкого
строительства
Ричмонд, Штат Калифорния 94804

Реферат

Сейсмоизоляция имеет место в настоящее время в самом рабочем состоянии разработки. Много новых типов систем изоляции исследуются и эластомерные изоляторы, система которая была использована на почти всех системах изоляции, законченных до настоящего времени продолжают подвергаться улучшению. По крайней мере одни дюжина больших проектов, или новый или модификация существующих зданий, были закончены и расчетные исследования находятся в стадии исследования по крайней мере другие одни дюжина больших проектов.

другие одни дюжина больших проектов. Большой экспериментальный исследовательский проект для изоляторов с применением ядерного реактора был выполнен в течение прошлого небольшого количества лет в [при] EERC. Эта программа влекла за собой испытание на сейсмостенде и испытание натуральных размеров и типовых изоляторов. Широкое разнообразие изоляторов были проверены включая низкий коэффициент формы, с умеренным коэффициентом формы, и очень с высоким коэффициентом формы эластомерные опоры.

Диапазон эластомеров которые были проверены включают низкое демпфирование, высокодемпфирующие, и очень низкомодульные составы. Полноразмерные и типовые изоляторы были проверены к отказу в нескольких характерах разрушения и коэффициенты безопасности для систем изоляции установились. Результаты испытаний показали который правильно разработали и произвели изоляторы для применений ядерного реактора может выдерживать [подтверждать] уровни нагрузки вне любого возможного входного сейсмического воздействия и показывать что отказ системы изоляции не может происходить перед отказом изолированного сооружения. Таким образом, использование изоляции может только иметь благотворные вклады в защиту атомных электростанций, внутренних трубопроводов, и оборудования.

Представление будет делать обзор последней разработки в выполнении изоляции фундамента и описывать результаты программы испытаний для соответствующего применения на атомные электростанции

English Machine Translation of Russian Machine Translation Results

Present state of Technology of Seismic isolation In United States of America

James M. Kelly
Professor Of building part
State university California in [at] Berkeley
Research centre of earthquake-resistive construction
Richmond, Staff California 94804

Abstract

Seismic isolation takes place at present in the most development operational condition. Many new types of systems of isolation there are investigated and elastomeric isolators, system of which had been used on nearly all systems of isolation, completed continue to be subject to date to improvement. At least one dozens of large designs, or new or modification of existing buildings, had been completed and design investigations are to stage of investigation at least other one dozen of large designs.

Large experimental research design for isolators using nuclear reactor had been performed to past small quantity of years in [at] EERC. This program entailed shake table test and testing of natural dimensions and type isolators. Wide variety of isolators had been tested including low shape factor, with moderate shape factor, and very with high shape factor elastomeric supports.

Range of elastomers of which had been tested include low damping, high-damping, and very low-modulus compositions. Fullsize and type isolators had been tested to failure in several failure mode and margins of safety for systems of isolation were installed. Results of testings indicated which developed properly

and produced isolators for applications of nuclear reactor can withstand [to confirm] levels of load outside of any possible seismic input and to indicate that failure of system of isolation be unable to take place before [in front of] isolated building failure. Thus, use of isolation can have only beneficial contributions to protection of internal pipelines, and equipment.

Submission will review last development in performance of base isolation and to describe results of test schedule for corresponding application.