Syntactic Parameters in the Phrasal Machine Translation

Elena Kozerenko, Alexander Khoroshilov, Alexei A. Khoroshilov
Institute of Informatics Problems of the Russian Academy of Sciences, Moscow, Russia

Abstract - *The paper deals with the methods and techniques of representing syntactic parse rules within the system of parameters of the example-based machine translation framework largely based on automatically compiled set-phrase dictionaries and translation memory. The syntactic structures are introduced into the vocabulary entries of the machine translation system.*

Keywords: machine translation, syntax, semantics, set-phrase dictionaries, machine learning, translation memory

1 Introduction

The main objective of the research and development presented in the paper is establishing the mechanisms for syntactic parsing in the systems based on previously translated texts, i.e. translation memory and example-based machine translation.

The concept of translation memory (sentence memory) appeared as an alternative to traditional machine translation. That concept can be regarded as an attempt to realize the idea of the Japanese computer scientist Makoto Nagao, that in the process of machine translation it is necessary to use the large corpora of parallel texts, earlier translated by humans. A more adequate approach is based on the concept of statistical machine translation (statistical-based machine translation), which is defined by some authors as a "sort of machine translation of texts, based on comparing of large corpora of language pairs". In contrast to traditional machine translation, statistical approach is based on statistical computation of matching probability and does not use the linguistic algorithms. Large corpora of parallel texts are necessary for operation of this system. A statistical mechanism of text analysis is used in the process of translation. This mechanism allows to select the variant for the word combination translation based on matching frequency of the language pair elements. A weak point of statistical systems is partial or total absence of a mechanism of grammatical rules analysis for source and target languages. Therefore, it is hard to imagine that the system, which does not analyze the text from the point of view of grammar, is able to release the correct translation of semantically complex texts. The systems of machine translation of texts simulate operation of a human translator. Their efficiency depends on how the nature of language operation and cognition are taken into account, and this nature has not yet been adequately studied. Therefore, the developers of machine translation must take into account the experience in international communication and translation activity, that was accumulated by mankind.

The starting point of machine translation was marked by the word-wise approaches. In this case, single words were considered to be basic units of sense expressing the concepts, and the sense of larger speech units (word combinations, phrases and utterance-length units) was supposed to be determined on the base of the sense of words comprising them. In dictionaries the use of word combinations along with single words was also admitted. But these combinations were mainly the idiomatic expressions, and their amount in dictionaries of machine translation systems was negligible in comparison with the amount of single words.That experience testifies that in the process of text translation the phraseological word combinations expressing the concepts rather than single words are the basic units of sense. The concepts are the elementary intellectual images, by the use of which it is possible to create more complex intellectual images corresponding to translated text.

2 Methods and limitations of the set-phrase machine translation and statistical approaches

The main thesis of the set-phrase translation is a statement that the concept names in texts are determined by word combinations rather than single words [1-6]. The meaning of units of higher level cannot be fully reduced to the sum of meanings of lower level units comprising them.
Set-phrase translation employs single-step compiling of bilingual frequency dictionaries of words and set-phrase word combinations. The orientation at the semantic-syntactic and mainly word-by-word translation alone could not lead to the solution of the basic problems of machine translation, because within language and speech the sense of units of higher level, as a rule, cannot be reduced or fully reduced to the sense of the lower level units comprising them. Almost all known systems related to traditional machine translation systems, developed in that direction. Later the developers of traditional systems began to include more terminological word combinations into their dictionaries.
A weak point of statistical systems is partial or total absence of a mechanism of grammatical rules analysis for source and target languages. A system which does not analyze the text from the point of view of grammar is unable to release the correct translation of semantically complex texts. Statistical machine translation allows to select the variant for the word combination translation based on matching frequency of the language pair elements. The main thesis of the set-phrase translation is a statement that the concept names in texts are determined by word combinations rather than single words, which is not always true. Set-phrase translation is mainly founded on a single-step compiling of bilingual frequency dictionaries of words and set-phrase word combinations.

The idea to create machine translation systems on the base of previously translated texts can be realized in different ways. The *first way* provides text translation with the use of statistic analysis of large corpora of bilingual texts in the process of translation. This way is known as "statistical machine translation" [7-14, 17,18].

In statistical machine translation (SMT) the task of translating from one natural language into another is treated as a machine learning problem. This means that via training on a very large number of hand-made translation samples the SMT algorithms master the rules of translation automatically. The first SMT developments were presented in [7,8]. The existing methods basically employ either sentence alignment or word alignment some experiments are made with phrase alignment and recently a mixed sentence-word approach has been developed to explore the paraphrases in the aligned parallel corpora. These attempts to consider linguistic information mark a step forward to acknowledging the intricate character of natural language if compared with other types of data. The mixed approach employs both sentence and word alignments. However, all these methods deal with the structural elements without considering the semantic aspects of the aligned language units.

The *second way* is connected with a single-step compiling of bilingual frequency dictionaries of words and phraseological word combinations. The creators of the systems of phraseological machine translations follow the second way [1-6]. This way excludes the fatal dependence of the translation process on availability of large volumes of parallel texts and quality of their translation.

Since the systems of phraseological machine translation are based on the theoretical concept, the main thesis of which is a statement that the concept names in texts are determined by word combinations rather than single words. Therefore, in the process of text translation from one language into another, it is necessary to use the phraseological combinations expressing the concepts, relationships between concepts and the typical situations rather than single words as basic units of sense. The single words may also be used if the translation with the help of the phraseological word combinations fails.

In compliance with this thesis, the system of phraseological machine translation must comprise the knowledge base of translation equivalents for most frequent phrases, phraseological combinations and single words. In the process of text translation the system should use the translation equivalents stored in its knowledge base in the following order: at first, an attempt to translate the successive sentence of the source text as the integral phraseological unit is made; then, if this attempt fails, words combinations being a part of the sentence should be translated; and, finally, if both above-mentioned attempts fail, word-by-word translation of the text fragments is performed. The fragments of the target text translated with the use of all three approaches, must grammatically agree with one another with the help of procedures of morphological and syntactic synthesis). Let us give consideration to this concept in detail. It is necessary to apply the following principles in the process of development of phraseological machine translation systems:

1. The phraseological units (word combinations and phrases) are basic language and speech units, which should be primarily included in the computerized dictionary.

2. Along with the phraseological units composed of continual word sequences, so called "speech models" - phraseological units with blank spaces, that may be filled with different words and word combinations, generating meaningful segments of speech, may be used in machine translation systems.

3. Real texts, without regard to their subject area, tend to be polythematic, if they have sufficiently large size. These texts differ from each other not so much by word stock as by probability distribution of occurrence of different words and word combinations from national word stock in them. Therefore, the computerized dictionary designed for translation of the text belonging to a single subject area must be polythematic, not to speak of translation of texts belonging to different subject areas.

4. Systems of phraseological translation need high-volume computerized dictionaries. Such dictionaries should be created on the base of computer-aided processing of parallel texts - bilingual texts, which are translations of each other, and in the process of translation system operation .

5. Along with the main high-volume polythematic dictionary, it is also reasonable to use a set of additional small-volume highly specialized dictionaries in systems of phraseological machine translation. The additional dictionaries should only contain information missing from the main dictionary (for example, data on priority translation equivalents of word combinations and words for different subject areas, if these equivalents are not equal to priority translation equivalents of the main dictionary).

6. The main means for solution of the problem of words polysemy in phraseological translation systems is their use in phraseological word combinations. The additional means is a set of additional specialized dictionaries , where the priority translation equivalent specific for subject area in question is identified for each multiple-meaning word or word combination.

7. The procedures of morphological and syntactic analysis and synthesis of texts, that are built on the base of linguistic analogy may play a major role in the systems of phraseological machine translation of texts.. These procedures allow to give up storing large amounts of grammatical information in dictionaries and generate it automatically in the process of translation when the need arises. They make the translation system open and capable of processing the texts with "new" words.

Along with text translation in automatic mode, it is reasonable to provide for an interactive mode of operation for the systems of phraseological machine translation. In that mode the user should have potentiality to intervene in the translation process and adapt the additional computerized dictionaries for the subject area of the translated text.

The phrase-based translation model, or the alignment template model and other similar approaches have greatly advanced the development of machine translation technology due to the extension of the basic translation units from words to phrases, i.e. the substrings of arbitrary size.

However, the phrases of the statistical machine translation model are not the phrases in the meaning of any existing syntax theory or grammar formalism, thus, for example, a phrase can be like “alignments the”, etc. A real challenge is the cross-level (e.g. morphology-to-syntax) matching of language structures in parallel texts. New research and development results demonstrate the growing awareness of the demand for enhancing linguistic motivation in statistical translation models and machine learning techniques [17,18].

3 Adult learning memory metaphor

In contrast to the approaches on the basis of “translation memory” that provide the increase of a machine translation system language competence by accumulating the previously translated text fragments and mainly based on regular expressions, Cognitive Transfer Grammar - CTG [15,16] is intended for the realization of the mechanism of structural memory, which simulates language competence of an adult learner (“Adult Learning Memory”). Thus, structural memory comprises the following components:

1) The initial basic collection of grammar rules represented in the formalized form (CTG);

2) The mechanisms of expansion and refinement of the system of rules, implemented by means of the methods of machine learning on parallel texts.

Our studies are based on the concepts of the functional approach, which we have used for the multilingual situation. With the development of the linguistic processor, which ensures English - Russian and Russian - English transfer, we introduced the concept of functional transfer fields (FTF) [16] that served the basis for the segmentation of language structures for the solution of machine translation problems. The basic idea of FTF consists in the adoption of the hypothesis about the fact that at the basis of grammatical structures there lie the cognitive structures (mental frames); a functional transfer field reflects the interaction of elements from different language levels.

“Adult learning memory” (ALM) means the employment of the “adult rule kit” a starter set of about 300 rules stating the structural semantic correspondences between source and target languages.

The machine translation technique comprises analysis, transfer and generation across the functional – categorial values of language units [15,16].

The process of structural patterns recognition is performed basing on the multiple transfer rule set and the probabilistic functional tree substitution grammar.

Translation activity involves the search for equivalence between structures of different languages. However, to establish whether the structures and units are equal or not, we need some general equivalent against which the language phenomena would be matched. Our approach based on the principle “from the meaning to the form” focusing on Functional Syntax would yield the necessary basis for equivalence search.

Consider some statistically relevant examples. Sometimes, a word may be translated by a word of another part-of-speech in the target language, a word combination, or even a clause, as the English word *implementable* is best translated into Russian as *kotoryi vozmozhno realizovat* (*which can be implemented*). To overcome these differences the categorial and functional features of the two languages were considered, and the structures of the input were made conformed to the rules of the target language by applying contrastive linguistic knowledge for implementation of the transfer model. A suitable formalism is indispensable for an algorithmic presentation of the established language transfer rules, and the language of Cognitive Transfer Structures (CTS) was developed based on rational mechanisms for language structures generation and feature unification [15].

4 Syntactic rules in digital dictionaries

First of all, the systems of phraseological machine translation should be aimed at translation of texts on business, science, technologies, politics and economy. Translation of fiction, artistic prose and poetic texts is a more complex and challenging task. But success can be also achieved in this area in future, if modern technological means are used to compile huge phraseological dictionaries for these texts.

Set-phrase units (word combinations and phrases) are basic language and speech units which should be primarily included in the computerized dictionary.

Along with the set-phrase units composed of continual word sequences, so called "speech models" - set-phrase units with slots that may be filled with different words and word combinations, generating meaningful segments of speech, can be used in machine translation systems.

At first glance, the machine translation concept offered by professor Makoto Nagao in 1984, fundamentally differs from the concept, formulated by professor G. G. Belonogov nine years earlier. But this is not true. Indeed, in the process of practical realization of Makoto Nagao’s concept, it is difficult to imagine that the text written in any language is completely the same as another text written earlier and translated into foreign language. It is not to be expected that this text contains long fragments (chapters, paragraphs and etc.), that are the same as the fragments of the text written and translated earlier. But, as our investigations showed, the continuous texts fragments including over ten words repeat on rare occasions - their total frequency doesn’t exceed 1%. It is necessary to use only short sentences, single words and text fragments (word combinations) including less than 10-12 words. This is the semantic-syntactic phraseological translation.

Of course, along with the translation equivalents of the relatively short fragments of texts, it is possible to include the translation equivalents of longer fragments in the computerized dictionaries. But in this case one should keep in mind, that the computerized dictionaries will be filled with "dead" ballast, i.e. with the dictionary entries, which will be used on rare occasions or will not be used at all in the process of text translation.

When developing the systems of phraseological machine translation, the most difficult and time-consuming problem appears to be the on of compiling sufficiently high-volume computerized dictionaries. The quality of translation depends on the volume of these dictionaries and on the quantity of the phraseological word combinations in them. And those volumes needs to be sufficiently large to provide the good covering of texts.

It is known, that in modern languages of the world (for example, in Russian or English) the amount of different words exceeds one million, and the amount of concept names determined by word combinations exceeds hundreds of millions. The authors of this article came to this conclusion on the basis of many years' experience of the statistical study of texts. Confirmation of such viewpoint is the report of the All-

European terminological centre "Infoterm" (Vienna, Austria, 1998), in which it was found that in modern languages of the world, such as English and German, a total amount of different terms exceeds 50 million, and nomenclature of goods exceeds 100 million. It is well known, that the connected texts consist of not only terms and names of goods.

The computerized dictionaries of such volume cannot be created quickly, but as experience shows, it is possible to achieve satisfactory quality of translation at the first stage in the presence of only several million entries in dictionaries, at least 80% of which should be word combinations. In this case the polithematic texts have the coverage of about 99, 7%.

Thereafter, the volume of dictionaries must be constantly increased and with the growth in amount of phraseological combinations, the quality of machine translation should improve. This problem cannot be solved by manual methods. For its solution, a system of computerized compiling and maintenance of the computerized dictionaries was created.

5 Semantic-syntactic set-phrase translation

It is necessary to use only short sentences, single words and text fragments (word combinations) including less than 10-12 words .

Along with the main high-volume polythematic dictionary, it is also reasonable to use a set of additional small-volume highly specialized dictionaries

The procedures of morphological and syntactic analysis and synthesis of texts, that are built on the base of *linguistic analogy* may play a major role in the systems of set-phrase machine translation of texts.

The implementation of the computerized phraseological text translation from one language into another must have three stages. At the first stage, the semantic-syntactic analysis of the source text is carried out. During that analysis the text is split into sentences, and then their conceptual and syntactic structure is determined. At the second stage (at the transfer stage) the concept names of the source text are substituted by the concept names in target language and the information on the syntactic structure of the source text is transformed into information required for the target text synthesis. At the final stage (the stage of semantic-syntactic synthesis of the target text) the text in the target language is formed.

The stages listed above are present in the process of translation of texts from any language to any other language, but their particular content for different pairs of languages has a specific character. This specific character can be seen in procedures of semantic-syntactic analysis and synthesis of texts, which include the procedures of morphological, syntactic and conceptual analysis and synthesis [1-6].

The set-phrase machine translation based on the multilingual dictionaries will operate in the same way, but these systems should be complemented with the procedures of semantic-syntactic and conceptual analysis and synthesis of all languages, which will be included in the system. The authors of this paper developed the effective technology based on the use of principles of linguistic analogy for creation of these procedures.

The computerized dictionaries are the most important part of the systems of phraseological machine translation. They should have sufficiently large volume, in order to cover texts, and should contain mainly word combinations. The authors developed the original methods, algorithms and programs for automated compiling and maintaining dictionaries for the system of phraseological machine translation. In cooperation with other specialists, the large-volume Russian-English and English-Russian phraseological computerized dictionaries containing 2, 6 million dictionary entries each were compiled. These dictionaries cover 99, 7% of the lexical content of modern texts and they represent the - powerful bilingual conceptual model for a wide range of fields of human activity. Phrase structures rules are incorporated into the vocabulary entries.

6 Well-formed nonterminals and dynamic rules

The main method for syntactic model enhancement in the set-phrase machine translation is including well-formed non-terminals in the general system of sentence analysis.

The non-terminals constitute the complete parse tree of a sentence comprising set-phrase models.

The dynamic formation of syntactic structures is supported by alternative categorial grammar parse on the basis of the rules dynamically extracted from parallel corpora.

Actually the process of transfer goes across the functional – categorial values of language units. A language structure which can be subjected to transfer has to be semantically complete from the point of view of its function. The cases of categorial shifts, in particular, when the technique of conversion is employed, require special treatment: the categorial shift of a syntax unit is determined by the functional role of this unit in a sentence (e.g. noun as a modifier is transformed into adjective).

The experience in creation of the large-volume Russian-English and English-Russian computerized dictionaries convinced the authors that Russian and English texts which are translations tino each other (for example, bilingual titles of the documents), can serve as the most reliable source for dictionaries compiling.

The compiling of the computerized dictionaries with the use of bilingual texts was carried out both manually and with the assistance of computers. The manual dictionary making requires huge expenditures of human labour. Therefore the authors of the article developed the procedure for automated dictionary making [3]. This procedure is based on the hypothesis, that in numerous bilingual pairs of sentences, which are translations of each other and which contain the same word or word combination of one of the languages, the word or word combination of another language, which is the translation of this word or word combination has maximal frequency. of occurence.

The procedure was used for processing bilingual (Russian and English) titles of the documents from the databases of VINITI (All-Union Scientific and Technical

Information Institute). In this case more than one million pairs of the document titles were processed. The computerized dictionaries of MetaPhrase system can be corrected and completed in the process of text translation in the interactive mode. In that mode there is an opportunity to identify the words and word combinations, which have no translation equivalents in the dictionary or these equivalents do not comply with the context or several equivalents are given, but the first equivalent does not comply with the context. These equivalents can be replaced by the equivalents complying with the textual context.

7 System performance

In compliance with the method described above, the large-scale experiment on compiling English-Russian frequency dictionaries on the base of the automated concept analysis of English and Russian titles of the documents, which are translations to each other, was carried out. For this purpose, the corpus of English titles of the polythematic documents and their Russian translations having the volume of about 2 million pairs of sentences from the VINITI's databases (1994-1999) were processed. The total volume of the corpus of texts is 390 Mb.

In the process of research three English-Russian frequency dictionaries were created:

1) the dictionary comprising the items which are the combinations of fragments of English and Russian titles of documents between which the translation equivalents were determined with the assistance of MetaPhrase system;

2) the dictionary comprising the items which are the fragments of titles of documents between which the translation equivalents have not been determined, but they are surrounded by the other fragments, between which such equivalents have been determined or by the signs of the beginning or the end of the title;

3) the dictionary comprising the items which are fragments of the English and Russian titles of documents between which the translation equivalents have been determined on the initial stage of titles processing.

The first frequency dictionary includes bilingual phraseological word combinations containing 2 to 16 words. It had 3.127.363 dictionary entries.

The value of the dictionary in question is that it contains translation equivalents between English and Russian fragments of titles of documents, which are longer than their fragments selected at the first stage of conceptual analysis of the titles. Each of the newly formed dictionary entries practically has just one translation version of an English word combination (the percentage of dictionary entries having more than one version of translation is less than 0.1).

The second frequency dictionary includes translation equivalents between fragments of titles of documents, that were not found at the initial stage of conceptual analysis of these titles. This dictionary contains 1.825.612 dictionary entries. The most frequent dictionary entries have the frequency of 1.008, and infrequent dictionary entries have- the frequency equal to one. 87% of dictionary entries have the frequency equal to one. A spot-check of the dictionary showed that about 50% of translation equivalents were incorrect. The quantity of such translation equivalents can be reduced at the final stage of dictionary making, if the procedure of semantic-syntactic checking is applied. After that, the dictionary must be edited by humans.

The third frequency dictionary contains translation equivalents between fragments of English and Russian titles of documents that were found at the initial stage of processing of these titles. It contains 387.025 dictionary entries. The most frequent dictionary entries have the frequency of 4.985, and most infrequent dictionary entries have- the frequency equal to one. 56% of dictionary entries had the frequency equal to one.

8 Digital set-phrase dictionaries

The authors developed the original methods, algorithms and programs for automated compiling and maintaining of the dictionaries for the system of set-phrase machine translation. Russian-English and English-Russian set-phrase dictionaries containing 2, 6 million dictionary entries each have been compiled. These dictionaries cover 99, 7% of the lexical content of modern texts and they represent powerful bilingual conceptual model for a wide range of fields of human activity.The compiling of the digital dictionaries with the use of bilingual texts was carried out both manually and with the assistance of computers. The procedure for automated dictionary making has been developed.

This procedure is based on the hypothesis, that in numerous bilingual pairs of sentences which are translations of each other and which contain the same word or word combination the translation of this word or word combination has maximal occurrence frequency.

The process of establishing concept equivalents is organized in the following way. At first stage, the semantic-syntactic analysis of the source text is carried out: the text is split into sentences and then their conceptual and syntactic structure is determined.

At the second stage (transfer) the concept names of the source text are substituted by the concept names in target language and the information on the syntactic structure of the source text is transformed into information required for the target text synthesis.

At the final stage (semantic-syntactic synthesis of the target text) the text in the target language is formed.

A system which does not analyze the text from the point of view of grammar is unable to release the correct translation of semantically complex texts.

Therefore, it proved to be necessary to introduce the well-formed non-terminals and parse rules into the representation mechanism of the set-phrase translation. It extended the compiling procedure of the dictionary as a means for automated linguistic processing.

9 Conclusion

In conclusion, it should be noted that this article describes the experience in creation of modern multilingual machine translation systems – the systems of phraseological translation. The extensive application of means of automation allowed to essentially reduce expenditures of human labour in the process of creation of this system, and therefore, to reduce the creation cost of such systems.

The modern multilingual machine translation systems should be based on set-phrase translation enhanced by cognitively and functionally motivated grammar.

The extensive application of means for automation allowed the authors to essentially reduce expenditures of human labour in the process of creation of the machine translation system, and therefore, to reduce the total creation cost of such systems.

Further research and development is connected with the dictionaries expansion and semantic-syntactic structuring.

10 Acknowledgements

The work presented in the paper was partially supported by the Russian Foundation for Basic Research, grant 11-06-00476-a.

11 References

[1] Belonogov, G. G., Khoroshilov, Alexander A., Khoroshilov, Alexei. A. Phraseological Machine Translation of Texts from Natural Languages to Other Natural Languages. . Col :"Scientific-Technical Information", Series 2. - M.: VINITI, 2010, № 10.

[2] Belonogov, G. G., Khoroshilov, Alexander A., Khoroshilov, Alexei A. Automatization of compiling of English-Russian bilingual phraseological dictionary using the corpora of bilingual texts. Col.: "Scientific Technical Information", Series 2. - M , VINITI, 2010, № 5.

[3] Belonogov, G. G., Gilyarevskij, R. S., Khoroshilov, A.A. On the nature of information. Col.:"Scientific Technical Information", Series 2. - 2009. - № 1.

[4] Belonogov, G. G., Kalinin, Yu. P., Khoroshilov, Alexander A., Khoroshilov, Alexei A.. Systems of Phraseological Machine Translation of Texts. Theoretical Preconditions and Experience in the Development. - M. 2007.

[5] Belonogov, G. G., Kalinin, Yu. P., Khoroshilov, A. A. Computational linguistics and Advanced Information Technologies. Theory and Practice of Constructing of Automatic Text Processing Systems. - M. 2004.

[6] Belonogov, G. G., Bystrov, I. I., Kozachuk, M V., Novoselov, A. P., Khoroshilov A.A. Automated Conceptual Text Analysis. Col. :" Scientific Technical Information", Series 2. - 2002. № 10.

[7] Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S. A statistical approach to machine translation // Computational Linguistics, 1990. Vol. 16. P. 79–85.

[8] Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L. The mathematics of statistical machine translation: Parameter estimation // Computational Lin- guistics, 1993. Vol. 19. No. 2. P. 263–311.

[9] Marino J. B., Banchs R. E., Crego J. M., de Gispert A., Lambert P., Fonollosa J. A.R., Costa-Jussa M. R. N-gram-based Machine Translation // Computational Linguistics, 2006. Vol. 32. No. 4. P. 527–549.

[10] Chen S. F. Aligning sentences in bilingual corpora using lexical information // Proceedings of the 31st Annual Conference of the Association for Computational Linguistics, 1993. P. 9–16.

[11] Callison-Burch C. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora // Proceedings of EMNLP-2008. 2008.

[12] Callison-Burch C., Koehn P., Monz C., Schroeder J. Findings of the 2009 Workshop on Statistical Machine Translation // Proceedings of Workshop on Statistical Machine Translation (WMT09), 2009.

[13] Och F. J., Ney H. The alignment template approach to statistical machine transla- tion // Computational Linguistics, 2004. Vol. 30. P. 417–449.

[14] Koehn P. and Hoang H. Factored translation models // Proceedings of the 2007Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007. P. 868–876.

[15] Kozerenko E. Features and Categories Design for the English-Russian Transfer Model // Advances in Natural Language Processing and Applications Research in Computing Science, 2008. Vol. 33. P. 123–138.

[16] Kozerenko E. B. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, Las Vegas, USA, 2003. — CSREA Press, 2003. P. 49–55.

[17] Wang W., May J., Knight K., and Marcu D. Re-Structuring, Re-Labeling, and Re- Aligning for Syntax-Based Statistical Machine Translation // Computational Lin- guistics, 36(2), 2010.

[18] Zhang H., Gildea D., and Chiang D. Extracting Synchronous Grammar Rules from Word-Level Alignments in Linear Time // Proceedings of the 22nd Interna- tional Conference on Computational Linguistics (COLING-08), Manchester, UK, August 2008.