ЛИНГВИСТИЧЕСКАЯ МОТИВИРОВКА ДЛЯ СТАТИСТИЧЕСКИХ МОДЕЛЕЙ ПЕРЕВОДА

Е. Б. Козеренко (kozerenko@mail.ru)

ИПИ РАН Москва, Россия

В данной статье рассматриваются проблемы выравнивания параллельных текстов для повышения достоверности перевода. Представлены статистическая и лингвистически-мотивированная модели выравнивания параллельных текстов и перевода методом трансфера. Предлагаемые решения основаны на гибридной грамматике, которая включает лингвистические правила и вероятностные характеристики структур языка. Поскольку сходные значения могут быть представлены различными способами, особенно важны описания синонимии языковых структур. Цель наших исследований — установление соответствий между структурами различных языков на уровне смысла.

Ключевые слова: перевод, параллельные тексты, выравнивание, модели выравнивания.

LINGUISTIC MOTIVATION FOR STATISTICAL TRANSLATION MODELS

E. B. Kozerenko (kozerenko@mail.ru)

Institute for Informatics Problems of the Russian Academy of Sciences, Moscow, Russian Federation

The paper deals with the problems of parallel texts alignment for enhancing the accuracy and adequacy of translation. Statistical and heuristic models of alignment and transfer are given. The solutions are proposed on the basis of a hybrid grammar, which includes linguistic rules and probabilities of language structures. The goal of the current development is the establishment of matches at the level of meaning, i. e. semantic matches. The meaning can be “packed” in different language structures, so the establishment of cross-language matches and inter-structural synonymy is of prime importance.

Key words: translation, parallel texts, alignment, alignment models.

1. Introduction

The paper is focused on discovering the ways of the two research paradigms combination, namely, introducing statistical methods into the rule-based systems of machine translation and employment of the methods and presentations capturing human language intuition in statistical translation models with the view of enhancing the existing language processing technologies.

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 [1,2].

The application of statistical models has considerably advanced the area of machine translation since the last decade of the previous century, however now new ideas and methods appear aimed at creating systems that efficiently combine symbolic and statistical approaches comprising different models. Both the paradigms move towards each other: more and more linguistics is being introduced into stochastic models of machine translation, and the rule-based systems include statistics into their linguistic rule systems. The procedures of analysis and translation are enhanced by the statistical data, which are taken into consideration by the "translation engine" for disambiguation of language structures. The stochastic approach to natural language processing originates from the projects in speech and characters recognition and spellcheckers. The main method for solving numerous problems, including the part of speech establishment and tagging, is the Bayesian approach. The architecture of stochastic systems is based on the dynamic programming algorithm.

Machine learning is rooted in the stochastic research paradigm. The training algorithms can be of the two types: supervised and unsupervised. An unsupervised algorithm should infer a model capable for generalization of the new data, and this inference should be based on the data alone. A supervised algorithm is trained on a set of correct responses to the data from the training set, so that the inferred model would provide more accurate decisions. The object of machine learning is the automatic inference of the model for some subject area basing on the data from this area. Thus a system learning, for example, syntactic rules should be supplied with a basic set of phrase structure rules. The widely used methods lately have been the N-grams which capture many intricacies of syntactic and semantic structures [3, 4], N-grams of variable length in particular [5], introduction of semantic information into N-grams. In [6] a detailed description is given of the approach to creating a statistical machine translation based on N-grams of bilingual units called "tuples" and the four special attribute functions.

The statistical models are built on the data obtained from the parallel corpora in different languages. Usually the texts are compared within language pairs. The text in the language from which the translation should be done is called the source text, and the text which is its translation is called the target text. Correspondently the languages are also called the source language and the target language (i. e. the language of translation).

The main method of extracting the data about the matches between the source and target languages and texts is the alignment of parallel texts. The result of this procedure is also called alignment and it is designated by A. The probability characteristics of alignments are employed in the algorithms of statistical machine translation. Hence, the alignment and the probability distribution are the key notions in these models description.

The following notations are employed in this paper: the symbol P denotes the probability distributions in the most general sense, and the symbol p denotes the probability distribution based on some particular model. The main attention in this paper is given to the description of various methods employed for parallel texts alignment, as the results of the alignment procedure determine the accuracy and adequacy of translation. We focus on the linguistic filters that are being introduced in the form of data structures and rules into the statistical translation models. The models under consideration are illustrated basing on the bilingual model for the Russian and English language pair. However, the similar methods are applicable for the alignments and translations of the Russian texts into the French and German languages, as well as other European languages.

2. Methods of parallel texts alignment

The statistical approaches to parallel texts alignment are aimed at establishing the most probable alignment A for the two given parallel texts S and T:

$$\arg\max_A P(A \mid S,T) = \arg\max_A P(A,S,T) \tag{1}$$

For estimation of the probability values indicated in this expression the most frequently used methods present the parallel texts in the form of aligned sentence sequences $(B_1, ..., B_K)$. The probability of each sequence is independent from the probabilities of other sequences, and it depends on the sentences in the given sequence only [7]. Then

$$P(A,S,T) \approx \prod_{k=1}^{K} P(B_k) \tag{2}$$

This method takes into account the length of sentences in the source language and in the target language measured in symbols. The longer sentence in one language will correspond to the longer sentence in the other language. This approach gives stable results for similar languages and literal translation. The more finely tuned mechanisms of matching are provided by the methods of lexical alignment. Thus in [8] the method of alignment by means of creating the model for consecutive word-by-word translation is presented. The best alignment result will be the one which maximizes the probability of a corpus generation with the given translation model. For the alignment of the two texts S and T they should be split into the sequences of sentence chains. A chain contains zero or more sentences in each of the two languages, and the sequence of chains covers the whole corpus

$$B_k = (S_{a_k}, ..., S_{b_k}; t_{c_k}, ..., t_{d_k}) \tag{3}$$

Then the most probable alignment $A = B_1, ..., B_{m_A}$ of the given corpus is determined by the following expression, and the chains of sentences do not depend on each other:

$$\arg\max_A P(S, T, A) = \arg\max_A P(L) \prod_{k=1}^{m_A} P(B_k), \quad (4)$$

where *P(L)* denotes the probability of the *L* chains being generated. The translation model employed in this approach is extremely simplified and does not take into account the factor of the word order in a sentence and the possibility of the fact that a word in the source text can correspond to more than one word in the text of translation. In this model the word chains are used, and they are limited to the 1:1, 0:1 *и* 1:0 matches. The essence of the model consists in the idea that if one word is usually translated by the word of another language, then the probability of the word chains matches 1:1 will be very high, and much higher than the product of probabilities of the 1:0 and 0:1 word chains matches where the given word occurs. And the program chooses the most probable alignment variant.

The translation model based on the word-by-word alignment (we employ this model for the Russian and English parallel texts) will be as follows:

$$P(r \mid e) = \frac{1}{Z} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(r_j \mid e_{a_j}), \quad (5)$$

where *e* is a sentence in English; *l* is the length of *e* expressed in words; *r* is a sentence in Russian; *m* is the length of *r*; r_j is the j-th word in *r*; a_j is the position in *e*, with which the r_j is aligned; $P(w_r \mid w_e)$ is the probability of translation, i. e. the probability of the w_r appearing in the Russian sentence if the corresponding w_e occurs in the English sentence, and *Z* is the normalization constant.

However, the above stated approach based on the word-by-word comparison and in no way accounting for the links between words and phrases does not give optimal results for the alignment of the Russian language and the English language texts, for there are certain structural differencies between these languages, and in translation there can be considerable transformations. If the languages under consideration are structurally different, the methods are used oriented at the introduction of grammar knowledge, for example, the alignment methods based on the words that belong to particular parts of speech [9] are employed. In this case the auxiliary words are not taken into account. For the employment of these methods the part of speech tagging of the parallel texts should be performed. The most general definition of the word-based alignment is given in [10]. Suppose the two word chains are given, one in the source text (for example, in Russian — r) $r_1^J = r_1, ..., r_j, ..., r_J$, and the other one is in the target language (English — e) $e_1^I = e_1, ..., e_i, ..., e_I$, and for these chains it is necessary to establish the alignment. The alignment between the two chains of words is a subset of a Cartesian product of the positions of words, i. e. the alignment A is defined as follows:

$$A \subseteq \{(j, i) : j = 1, ..., J; i = 1, ..., I\}. \quad (6)$$

In machine translation based on statistical methods an attempt is made to construct a model of the translation probability $P(r_1^J \mid e_1^I)$, which describes the correlation between some chain r_1^J in the source language and the chain e_1^I in the target language. In statistical texts alignment model $P(r_1^J, a_1^J \mid e_1^I)$ a "hidden" alignment a_1^J is introduced which describes the mapping from the source position j into the target position a_j. The correlation between the translation model and the alignment model is given in the following way:

$$P(r_1^J \mid e_1^I) = \sum_{a_1^J} P(r_1^J, a_1^J \mid e_1^I). \tag{7}$$

The alignment a_1^J can contain the alignments $a_j = 0$ with the empty word e_0 for the words of the source language which had not been aligned with any word in the source language. On the whole the statistical model depends on the set of unknown parameters θ which are extracted from the training data set in the course of learning. The following presentation is used to express the dependence of the model on the set of parameters:

$$P(r_1^J, a_1^J \mid e_1^I) = p_\theta(r_1^J, a_1^J \mid e_1^I) \tag{8}$$

The technique of statistical modeling consists in the development of specific statistical models which would capture the most relevant features of the subject area under consideration. Thus a statistical model of alignment should adequately describe the correlation between the chain in the source language and the chain in the target language.

For detection of the unknown parameters θ a training corpus of parallel texts is given containing S sentence pairs $\{(r_s, e_s) : s = 1, ..., S\}$. For each pair (r_s, e_s) the alignment variable is designated by $a = a_1^J$. The unknown parameters are established by means of maximization of the parallel texts similarity in the corpus:

$$\hat{\theta} = \arg\max_{\theta} \prod_{s=1}^{S} \sum_{\mathrm{a}} p_\theta(\mathrm{r}_s, \mathrm{a} \mid \mathrm{e}_s). \tag{9}$$

As a rule the maximization for such models is performed on the basis of the expectation maximization algorithm [11] or the similar ones. Such algorithm is useful for the solution of the parameters estimation problem, but it is not indispensable for the statistical approach.

Hence despite the fact that there exist a large number of alignments for a given pair of sentences, it is always possible to find the best alignment:

$$\hat{a}_1^J = \arg\max_{a_1^J} p_{\hat{\theta}}(r_1^J, a_1^J \mid e_1^I). \tag{10}$$

The alignment $\hat{a}_1^J$ is also called the Viterbi alignment for the pair of sentences (r_1^J, e_1^I). The estimation of the Viterbi alignment quality is performed by means of comparison with some reference alignment carried out manually. The parameters of statistical alignment models are optimized with the consideration of the maximal likelihood criterion which does not always reflect the quality of alignment.

The most frequently used statistical model which is used for parallel texts alignment is the hidden Markov model [13]. The alignment model $P(r_1^J, a_1^J \mid e_1^I)$ can be structured without the loss of generality in the following way:

$$P(r_1^J, a_1^J \mid e_1^I) = P(J \mid e_1^I) \cdot \prod_{j=1}^{J} P(r_j, a_j \mid r_1^{j-1}, a_1^{j-1}, e_1^I) = \\ = P(J \mid e_1^I) \cdot \prod_{j=1}^{J} P(a_j \mid r_1^{j-1}, a_1^{j-1}, e_1^I) \cdot P(r_j \mid r_1^{j-1}, a_1^j, e_1^I) \quad (11)$$

When using this alignment the three probabilities are obtained: a length probability $P(J \mid e_1^I)$, an alignment probability $P(a_j \mid r_1^{j-1}, a_1^{j-1}, e_1^I)$ and a lexicon probability $P(r_j \mid r_1^{j-1}, a_1^j, e_1^I)$. In the hidden Markov alignment model the first order dependence for the alignments a_j is assumed, and it is assumed that the lexicon probability depends only on the word at position a_j:

$$P(a_j \mid r_1^{j-1}, a_1^{j-1}, e_1^I) = p(a_j \mid a_{j-1}, I), \quad (12)$$

$$P(r_j \mid r_1^{j-1}, a_1^j, e_1^I) = p(r_j \mid e_{a_j}). \quad (13)$$

If a simple length model is assumed $P(J \mid e_1^I) = p(J \mid I)$, then for $p(r_1^J \mid e_1^I)$ the following decomposition based on the hidden Markov model is obtained:

$$p(r_1^J \mid e_1^I) = p(J \mid I) \cdot \sum_{a_1^J} \prod_{j=1}^{J} [p(a_{j-1}, I) \cdot p(r_j \mid e_{a_j})] \quad (14)$$

with the alignment probability $p(i \mid i', I)$ and the translation probability $p(r \mid e)$. In order to make the alignment parameters independent from the absolute values of word positions, it is assumed that the alignment probabilities $p(i \mid i', I)$ depend only on the jump width $(i - i')$. Using a set of non-negative parameters $\{c(i - i')\}$, it is possible to present the alignment probabilities in the following way:

$$p(i \mid i', I) = \frac{c(i - i')}{\sum_{i''=1}^{I} c(i'' - i')}. \quad (15)$$

This form ensures that the alignment probabilities satisfy the normalization constraint for each conditioning word position $i', i' = 1, ..., I$. This model is also called the homogeneous hidden Markov model [12]. The original formulation of the hidden Markov alignment model did not comprise the empty word generating source words which have no directly aligned word in the target text. In [13] the empty word is introduced and the hidden Markov model network is extended by means of I empty words e_{I+1}^{2I}.

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 [14, 15]. However, all these methods deal with

the structural elements without considering the semantic aspects of the aligned language units.

The phrase-based translation model, or the alignment template model [16] and other similar approaches have greatly advanced [17] 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 this 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 [18]. New research and development results demonstrate the growing awareness of the demand for enhancing linguistic motivation in statistical translation models and machine learning techniques [21,22].

3. Intertext development: establishment of semantic matches

The above stated methods are being employed for design and development of a linguistic knowledge base Intertext. It is a linguistic resource with semantic grouping of phrase structure patterns provided with the links to synonymous structures at all language levels for the languages included into the linguistic base.

Our focus on configurations provides high portability to the language processing software designed under these principles: we can operate with a lexicon which has only standard linguistic information including morphological characteristics, part of speech information and the indication of transitivity for verbs.

The Intertext linguistic knowledge base comprises the following components:

- parallel texts database: the texts are segmented into the functionally relevant structures that are semantically aligned;
- a bilingual Treebank (under development at present);
- structural parse editor (under development at present) which displays the parse and transfer schemes for indicated text segments;
- the inventory of structural configurations arranged on Cognitive Semantic principle.

4. Establishment of cross-language matches and inter-structural synonymy

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.

4.1. Types of matches

The following types of structural semantic matches have been observed:

word → word, phrase structure → phrase structure, word → phrase structure, morpheme → word, morpheme → phrase structure.

Syntactically languages are most different in the basic word order of verbs, subjects, and objects in declarative clauses. English is an SVO language, while Russian has a comparatively flexible word order. The syntactic distinction is connected with a semantic distinction in the way languages map underlying cognitive structures onto language patterns, which should be envisaged in MT implementations [20].

The basis of Cognitive Transfer Grammar (CTG) is composed of the proto-typical structures of the languages (in the initial model Russian and English) being investigated, their most probable positions in a sentence, statistical data about the distributive characteristics of structures (the information about the contextual conditions of the use of the investigated objects, i. e. the information about the structural contexts), the schemes of the complete parse of sentences.

The creation and development of the CTG assumes: — the semantic approach to the analysis of language meaning and language form (forms); — the construction of formal grammar presentations taking into account the structures of components and mechanisms of linearization, and also the relations of dependence between the units of a syntactic tree (the approach, which has the features of similarity to HPSG: the inheritance of the features via the head elements of phrase structures); — the inclusion of the probability characteristics of language objects; — the creation of Cognitive Transfer Spaces (CTS), represented in the form of expert linguistic rules, which can be extended by means of the establishment of synonymous language structures of parallel texts in different languages. The notion of Cognitive Transfer Spaces is the elaboration of the Functional Transfer Fields idea (see Section 5) for the multivariant translations of language structures.

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 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) [19] 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.

The basic design unit of the spaces of cognitive transfer is a *transfeme*.

Definition. *Transfeme* is a unit of cognitive transfer the, i. e. a semantic element embodied in a translatable semantically relevant language segment taken in the unity of its categorial and functional characteristics, that establishes the semantic correspondence between the language structures, which belong to different language levels and systems. The types of transfemes are determined by the rank of transfemes.

We distinguish the following ranks of transfemes:

- rank 1: lexemes as structural signs, i. e., a word, considered as a categorical — functional unit without taking into account the specific lexical value of this word;
- rank 2: a word combination, i. e., the syntactic structure, which consists of two and more syntactically connected words, but never a complete sentence (clause);
- rank 3: a clausal unit, i. e., dependent (subordinate) clause;
- rank 4: a sentence (either a simple sentence or the main clause of a complex sentence);
- rank 5: a scattered structure, i. e., a word group, which is characterized by a syntactic and semantic unity, but is discontinuous, i. e., between the members of the group there appear other language objects, which are not the members of this group;
- rank 0: the morphological units, which are not independent words, but which form a part of a lexeme of a source language, and in the language of transfer can be expressed by a clause and the units of other ranks, for example: the suffixes — *ible, — able* which are synonymous to the construction *"which can be"*, e. g. *extensible — which can be extended.*

4.2. Cross-level focus

Our studies focus on particular situations when the semantic match goes across language levels. The segmentation of phrase patterns used for the input language parse was carried out with the consideration of semantics to be reproduced via the target language means. Both the most important universals such as enumeration, comparison, modality patterns, etc., and less general structures were singled out and assigned corresponding target language equivalents.

Consider an example of a phrase structure conveying the modal meaning of obligation: "*…the task to be carried out…*". In other words, the meaning of this phrase can be rendered as "*…the task that should be carried out…*". The Infinitive phrase in the Engish language gives the regular way of expressive means compression without the loss of semantic value. A literary translation in Russian requires the second way of presenting the same idea of obligation. However in this specific case a "reduced" translation variant is also possible which consists in the introduction of the subordinate conjunction "*chtoby*" — "*so that*", between the noun and the modifying Infinitive. The parse rule would look like: *NP(to) → NP VPto; a*nd the generation rule would be presented as: *NP(to) → NP Punct.{comma} Conj.(chtoby) VPto*.

Special attention is required for the problem of passive constructions transfer. As in the phrase "*was considered*". The rules for simultaneous translation (which in many cases is similar to the real time machine translation performance and can be a source of compromise decisions for phrase structure design) requires the transformation of the English Subject into the Direct Object (Russian, Accusative Case) standing in the first position in a sentence and the passive verbal form would produce an impersonal verbal form in Russian.

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 → adjective).

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 *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.

We apply multivariant CTG constraints to our parse and transfer algoritm to choose the optimal variants for translations from English into Russian (and from Russian into English). Each phrase (transfeme) has a set of different CTG labels, and we need a way of choosing which label to use when applying the constraint. At present we choose the best label for the phrase in a parse tree and the best transfer variant in the language of translation:

$$e = \arg\max_{e} \arg\max_{s \in CTG-labels(e,P)} p(e \mid r, s) \tag{16}$$

where *e* is an Enlish sentence, *r* is a Russian sentence, *P* is an English parse tree, *s* is a syntactic type of e belonging to the Cognitive Transfer Grammar.

Our linguistic simulation efforts are aimed at capturing the cross-level synonymy of language means and cross-linguistic semantic configurational matches for the English and Russian languages. The emphasis on the practical human translation experience gives the reliable foundation for statistical studies of parallel text corpora and automated rule extraction in further studies.

5. Rule set for training data: cognitive semantic approach

The establishment of structures equivalence on the basis of functional semantics proved to be useful for developing the syntactic parse and transfer rules module

for the English — Russian machine translation. This rule module was implemented in the first release of the Cognitive Translator system [19,20]. Generally, major efforts connected with natural language modeling lay emphasis at lexical semantics presentations and less attention is paid to the semantics of structures and establishment of functional similarity of language patterns as a core problem in multilingual systems design.

The set of functional meanings together with their categorial embodiments serves the source of constraints for the unification mechanism in the formal presentation of our grammar. The formalism developed employs feature-based parse, and head-feature inheritance for phrase structures which are singled out on the basis of functional identity in the source and target languages. The transferability of phrase structures is conditioned by the choice of language units in the source and target languages belonging to the same functional transfer fields (FTF), notwithstanding the difference or coincidence of their traditional categorial values. A set of basic FTF was singled out and language patterns employed for conveying the functional meanings of interest were examined:

- Primary Predication FTF (non-inverted) bearing the Tense — Aspect — Voice features; this field mainly includes all possible complexes of finite verbal forms and tensed verbal phrase structures.
- Secondary Predication FTF bearing the features of verbal modifiers for the Primary Predication FTF. Included here are the non-finite verbal forms and constructions, and subordinate clauses comprising the finite verbal forms. All these are united by the functional meanings they convey, e. g. qualification, circumstance, taxis (ordering of actions), etc.
- Nomination and Relativity FTF: language structures performing the nominative functions (including the sentential units) comprise this field.
- Modality and Mood FTF: language means expressing modality, subjunctivity and conditionality are included here. Here the transfer goes across the regular grammatical forms and lexical means (modal verbs and word combinations) including phrasal units.
- Connectivity FTF: included here are lexical — syntactic means employed for concatenation of similar syntactic groups and subordination of syntactic structures.
- Attributiveness FTF: adjectives and adjectival phrases in all possible forms and degrees comprise the semantic backbone of this field; included here are also other nominal modifiers, such as nominative language units and structures (*stone wall* constructions, prepositional genitives — *of* –phrases), and other dispersed language means which are isofunctional to the backbone units.
- Metrics and Parameters FTF: this field comprises language means for presenting entities in terms of parameters and values, measures, numerical information.
- Partition FTF: included in this field are language units and phrase structures conveying partition and quantification (e. g. *some of, part of*, *each of*, etc.).
- Orientation FTF: this field comprises language means for rendering the meaning of space orientation (both static, and dynamic).
- Determination FTF: a very specific field which comprises the units and structures that perform the function of determiner (e. g. the Article, which is a good

example for grammar — lexical transfer from English into Russian, since in Russian there exist no such grammatical category; demonstrative pronouns, etc.).
- Existentiality FTF: language means based on *be*-group constructions and synonymous structures (e. g. sentential units with existential *there* and *it* as a subject: *there is…*; *there exists…*; etc.).
- Negation FTF: lexical — syntactic structures conveying negation (e. g. *nowhere to be seen*, etc.).
- Reflexivity FTF: this field is of specific character since the transfer of reflexivity meaning goes across lexical — syntactic — morphological levels.
- Emphasis — Interrogation FTF: language means comprising this field are grouped together since they employ grammar inversion in English.
- Dispersion FTF: individual language structures specific for a given language are included here; these are presented as phrasal templates which include constant and variable elements. We single out 3 major types of the Dispersion FTF.

Interpretation techniques employ the segmentation of structures carried out on the basis of the functional transfer principle. The principal criterion for including a language structure into a field is the possibility to convey the same functional meaning by another structure of the field, i. e. the interchangeability of language structures. A constraint-based formalism which is called the Multivariant Cognitive Transfer Grammar has been developed and. It comprises about 350 transferable phrase structures together with the multiple transfer rules combined within the same pattern. Such patterns, or Cognitive Transfer Structures (CTS), serve constitutional components of the declarative syntactical processor module and encode both linear precedence and dependency relations within phrase structures. Consider, for example, the functional meaning of *Possessiveness*, which belongs to the Functional Transfer Field of *Attributiveness* in the following phrases: *Peter's house; the house of Peter.*

However, we see our main objective not in creation of an abstract semantic meta language, but in a careful research of all possible kinds of configurations of language patterns used by natural languages for expression of functional meanings.

5.1. Linguistic filters on the basis of the Cognitive Transfer Grammar

The key idea of our linguistic framework is cognitive cross-linguistic study of what can be called *configurational* semantics, i. e. the systemic study of the language mechanisms of patterns production, and what meanings are conveyed by the established types of configurations. We explore the sets of meanings fixed in grammar systems of the languages under study. Our studies are focused on the types of meanings outside the scope of lexical semantics, and we consider the lexical semantics when the meanings which we denote as configurational, have expression at the lexical level. The importance of this aspect is connected with the fact that natural languages are selective as to the specific structures they employ to represent the referential situation. However, it is always possible to establish

configurations which perform the same function across different languages (i. e. isofunctional structures). The parse aimed at transfer procedures requires a semantic grammar and cannot be efficiently implemented through a combination of monolingual grammars.

In the previously formulated Cognitive Transfer Grammar (CTG) [19, 20] the functional meanings of language structures are determined by the categorial values of head elements. The probability characteristics are introduced into the rules of the unification grammar as weights assigned to the parse trees.

In the Cognitive Transfer Grammar the basic structures are the *transfemes*. A *transfeme* is a unit of cognitive transfer establishing the functional semantic correspondence between the structures of the source language L_S and the structures of the target language L_T. For the alignment of parallel texts the transfemes are given as the rewrite rules in which the left part is a nonterminal symbol, and the right part are the aligned pairs of chains of terminal and nonterminal symbols which belong to the source and target languages :

$$T \to \langle \rho, \alpha, \sim \rangle, \tag{17}$$

where T is a nonterminal symbol, ρ and α are chains on terminal and nonterminal symbols which belong to the Russian and English languages, and ~ is a symbol of correspondence between the nonterminal symbols occuring in ρ and the nonterminal symbols occuring in α. In the course of parallel texts alignment on the basis of the CTG the derivation process begins with a pair of the linked starting symbols S_r and S_α, then at each step the linked nonterminal symbols are rewritten pairwise with the use of the two components of a single rule.

5.2. CTG-alignment

For automatic extraction of the rules on the basis of CTG from parallel texts these texts should be previously aligned by sentences and words. The extracted rules base on the wordwise alignments in such a way that at first the the starting phrase pairs are identified with the use of the same criterion as the majority of statistical models of translation employing the phrase-based approach [16], which means that there should be at least one word inside a phrase in one language aligned with some word inside a phrase in another language, but no word inside a phrase in one language can be aligned with any word outside its pair phrase in another language.

Definition 1. Assume that a pair of sentences $\langle r, e, \sim \rangle$ aligned wordwise is given, assume that r_1^j denotes a substring r from the position i to the position j inclusive, and correspondingly, $e_{i'}^{j'}$ denotes a substring e from the position i' to the position j' inclusive. Then the rule $\langle r_1^j, e_{i'}^{j'}, \sim \rangle$ is a starting phrase pair.

In order to continue the extraction of rules from the singled out phrases we find the phrases which contain other phrases and substitute them by nonterminal symbols.

Thus the mechanism of rules embedding is implemented which reflects the hierarchical structure of the natural language.

The next step is the formation of the rule system in the CTG notation. Cognitive Transfer Grammar is a generative unification grammar having a hierarchical structure and reflecting a major part of language transformations employed in the process of translation from one language into another. Besides, basing on the experimental data obtained from the corpora study the CTG rules are supplied with the weights of possible derivation variants.

Definition 2. Cognitive Transfer Grammar G_{CT} is a set

$$G_{CT} = \{T_{L_1}, T_{L_2}, N_{L_1}, N_{L_2}, P_{CA}, P_{CT}, S_{L_1}, S_{L_2}, M, D\}, \quad (18)$$

Where T_{L_1}, T_{L_2} are the sets of terminal symbols of the languages L_1 and L_2; N_{L_1}, N_{L_2} are the sets of non-terminal symbols of the languages L_1 and L_2; P_{CA}, P_{CT} are the rules of analysis and synthesis on the basis of the cognitive transfer ; S_{L_1}, S_{L_2} are a pair of the starting symbols of the languages L_1 и L_2 with which the process of analysis and alignment of sentences is initiated; M is the function of establishing the correlations between the structures of the languages L_1 and L_2; D is the function assigning the probability values to each rule from the sets P_{CA}, P_{CT}.

Ambiguity is an immanent feature of the natural language and it is a cause of major difficulties in machine translation implementation. Ambiguous and polysemous syntactic structures are taken into account in the further development of the CTG mechanisms, which is the multivariant CTG, and the implementations of the multivariant CTG data structures are designed as linguistic filters for statistical translation models. These data structures are called multivariant cognitive transfer structures (MCTS). The general presentation of the MCTS syntax is as follows :

MCTS {MCTS <identifier> MCTS <weight> MCTS <tag>}→
<Input phrase structure and the set of its features and values > →
<Head-driven transfer scheme> →
<Generated phrase structure and its set of features and values — variant 1>
<weight 1>
<Generated phrase structure and its set of features and values — variant 2>
<weight 2>
<Generated phrase structure and its set of features and values — variant N>
<weight N> .

The new multivariant CTG captures the polysemy of syntactic structures, the mechanisms of disambiguation basing on statistical data are introduced into the systems of parse and transfer rules, possible contexts of language structures are taken into account.

The multivariant CTG provides an extensible platform for the development of machine translation and knowledge extraction systems. At present the CTG principles are employed for development of the rule systems for the Russian-French and

Russian-German language pairs. A new hybrid approach to construction of the models for machine translation and other natural language processing systems bridges the gap between symbolic and stochastic paradigms. The new training data sets are introduced into the linguistic knowledge base for upgrading the rule systems. The linguistic filters employed for reduction of the noise rules generated in the process of learning are based on the cognitive transfer spaces which comprise major groups of cross-lingual functional synonyms.

Conclusions The urgency of the new hybrid methods of language objects presentation is caused by the demand for the optimal combination of advantages of the two research paradigms: logical linguistic modelling employing the designed rules and stochastic approach based on machine learning. This development is of special importance for the tasks of structural analysis and computer modelling of the full text scientific and patent documents. The work with patent documents requires the introduction of specific features of patent texts: such as employment of certain language constructions, the syntax of patent formulae, the extensive use of templates, domain-oriented lexicons. The Intertext base comprises a collection of scientific and patent texts in the Russian and English languages from the areas of Computer Science, Social Monitoring, Chemical Technology and other areas. One of the latest developments is connected with implementing the natural language web service for the multilingual search and analysis of financial information.

The objectives of the prospective research and development efforts consist in the inclusion of parallel texts and language processing features for the French, German and Italian languages, and evolving the Intertext into a multilingual knowledge base. Our focus on configurations provides high portability to the language processing software designed under these principles: we can operate with a lexicon which has only standard linguistic information including morphological characteristics, part of speech information and the indication of transitivity for verbs. The approach taken would be important in further development of educational programs for computer science and computational linguistics courses. Educational relevance of the methods discussed in the paper lies in deeper understanding of uniform cognitive mechanisms employed in particular language embodiments of semantic structures.

References

1. *Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S.* 1990. A Statistical Approach to Machine Translation. Computational Linguistics,16 : 79–85.
2. *Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L.* 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19 (2) : 263–311.
3. *Callison-Burch C.* 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. Proceedings of EMNLP-2008.

4. *Chen S. F.* 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. Proceedings of the 31st Annual Conference of the Association for Computational Linguistics : 9–16.
5. *Dempster A. P., Laird N. M., Rubin D. B.* 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Ser. B, 39 (1) : 1–22.
6. *Gale W. A., Church K. W.* 1993. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19 : 75–102.
7. *Niesler T. R., Woodland P. C.* 1999. Modelling Word-Pair Relations in a Category-based Language Model. IEEE ICASSP-99, IEEE : 795–798.
8. *Ney H., Essen U., Kneser R.* 1994. On Structuring Probabilistic Dependencies in Stochastic Language Modeling. Computer Speech and Language, 8 : 1–38.
9. *Marino J. B., Banchs R. E., Crego J. M., de Gispert A., Lambert P., Fonollosa J. A. R., Costa-Jussa M. R.* 2006. N-gram-based Machine Translation. Computational Linguistics, 32 (4) : 527–549.
10. *Masahiko H., Yamazaki T.* 1996. High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information. ACL 34 : 131–138.
11. *Och F. J., Ney H.* 2000. A Comparison of Alignment Models for Statistical Machine TransLation. COLING'00: The 18th International Conference on Computational Linguistics : 1086–1090.
12. *Och F. J., Ney H.* 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29 (1) : 19–51.
13. *Rosenfeld R.* 1996. A Maximum Entropy Approach to Adaptive Statistical Language Modeling. Computer Speech and Language, 10 : 187–228.
14. *Vogel S., Ney H., Tillmann Ch.* 1996. HMM-based Word Alignment in Statistical Translation. COLING'96: The 16th International Conference on Computational Linguistics : 836–841.
15. *Callison-Burch C., Koehn P., Monz C., Schroeder J.* 2009. Findings of the 2009 Workshop on Statistical Machine Translation. Proceedings of Workshop on Statistical Machine Translation (WMT09).
16. *Och F. J., Ney H.* 2004. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, 30 : 417–449.
17. *Koehn P., Hoang H.* 2007. Factored Translation Models. Proceedings of the 2007Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) : 868–876.
18. *Yeniterzi R., Oflazer K.* 2010. Syntax-to-Morphology Mapping in Factored PhraseBased Statistical Machine Translation from English to Turkish. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics : 454–464.
19. *Kozerenko E. B.* 2003. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms. Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications : 49–55.
20. *Kozerenko E.* 2008. Features and Categories Design for the English-Russian Transfer Model. Advances in Natural Language Processing and Applications Research in Computing Science, 33 : 123–138.

21. *Wang W., May J., Knight K., Marcu D.* 2010. Re-Structuring, Re-Labeling, and ReAligning for Syntax-Based Statistical Machine Translation. Computational Linguistics, 36(2).
22. *Zhang H., Gildea D., Chiang D.* 2008. Extracting Synchronous Grammar Rules from Word-Level Alignments in Linear Time. Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08).