OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 681.3.06

COGNITIVE LINGUISTIC PRESENTATIONS OF LANGUAGE STRUCTURES IN ARTIFICIAL INTELLIGENCE AND MACHINE TRANSLATION SYSTEMS

Kozerenko E.B.

Institute of Informatics Problemsof the Russian Academy of Sciences

kozerenko@mail.ru

The paper focuses on the issues of establishing semantic content of syntactic structures in the English and Russian languages for the tasks of machine translation and knowledge management. The problem of establishing transferable language phrase structures is considered. The approach employed is based on generalized cognitive entities manifested in the categorial systems of a subset of natural languages (English and Russian in our case) and functional roles of language units in a sentence. A declarative module of syntactical processor was designed and implemented within the framework of machine translation system "Cognitive Translator" and a number of intelligent knowledge-based systems.

Keywords: machine translation, phrase structures, syntax, semantics, transfer.

INTRODUCTION

The present state of research and development in the field of machine translation and multilingual systems design requires new methods of linguistic reality presentations capturing the intricate features of natural languages and comprising the facilities of the already existing approaches. The crucial problem to be faced is *categorization* of linguistic phenomena. Of special concern are the syntactic-semantic structures since neither constituency grammar nor dependency grammar alone gives the complete expressive means for such natural language properties as syntactic ambiguity and synonymy.

Translation is a creative and sophisticated human activity, hence, producing automatically a high-quality translation of an arbitrary text from one language to another is a task too far from its complete implementation. However, for simpler tasks, such as acquiring information on the Web, getting acquainted with subject domain information, etc., rough translation output without post editing can be quite adequate. One of the domains where MT works best is scientific discourse.

Of the three forms of translation performed by man: written translation, consecutive interpretation and simultaneous interpretation, the one which is nearest to the real-time machine translation is simultaneous interpretation (SI). Therefore, the recommendations for SI are of prime interest to MT designers, as they propose more implementable solutions for lexical grammatical transformations than the first two forms.

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 [Nirenburg, et al.,1992]. Besides, there exist syntactic constructions specific of a given language (such as, for example, English constructions with existential "there" and "it" as formal subjects). Sometimes, a word may have translation to 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 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. The formalism developed for presentation of syntactic structures for the English-Russian machine translation is a variant of unification grammar and comprises over two hundred rules and it was implemented within the framework of machine translation system "Cognitive Translator" and a number of intelligent knowledge-based systems.

1. SI Techniques for Handling Syntactic Structures

Segmentation and unification of utterances in the course of translation is a major task for human professional interpreters. They would even say that syntax is “interpreter’s enemy”. The selectivity of languages as to the choice of specific characteristics of description of one and the same situation results in numerous distinctions, and one of the most crucial of them is the degree of particularity in conveying a referential situation. Therefore, a situation which in one language is described by means of one specific feature, in another language may require two or more characteristics. Thus, in many cases the English language is more economical (about thirty percent, according to the reports of simultaneous interpreters) [Visson, 1989, Visson, 1991] in expressing a thought than Russian.

A very good illustration of this phenomenon is attributive word combinations of the “stone wall” type which when being translated into Russian in many cases require numerous additions. On the other hand, Russian input in some cases may result in an expanded English translation.

In practice the technique applied to overcome this problem is *utterance segmentation* which consists in sectioning a source Russian sentence into two or more utterances in the resulting English sentence.

Another important rule is the least possible change of word order. But this inflicts other unavoidable transformations, and not all of them are implementable within the framework of machine translation. For example, the general rule for interpreters: a Russian noun which appears at the very beginning of a sentence and has the form of an oblique case, i.e. indirect object standing at the beginning of a Russian sentence, should be transformed into the subject of an English sentence notwithstanding its initial syntactic role

e.g. *Na vstreche dogovorilis’*…(*At the meeting agreed*…)

should be translated as -

The meeting reached an agreement…

This transformation performed in the course of human simultaneous interpretation appears to be unattainable to a machine translator at the present state of the art. The requirement of denotational equivalence involves numerous lexical grammatical shifts which cause transformations of the semantic structure of an utterance [Visson, 1989, Visson, 1991]. Another regular semantic shift, that of substituting a predicate of action by the predicate of state.

e.g. *He is a member of the college team*. (A predicate of state).

On igraet v studencheskoi komande. (He plays in the students’ team. A predicate of action).

Moreover, the existence of such shifts within the real text corpora inflicts complications for one more computational linguistics problem, that is text alignment, which in some cases may appear even intractable.

The following SI techniques appeared to be of use for MT design in the course of our development.

(1) Full translation of lexical grammatical forms is applied when these forms completely correspond to each other both in the source and the target languages as to their form, function and meaning.

(2) Null translation is applied when a grammatical form exists in the source and target languages but is used differently for explicating a certain referential situation.

(3) Partial translation is used when one and the same grammatical form has several content functions which differ in the source and target languages.

(4) Functional substitution is employed when the functions and meanings of grammatical forms in the source and target languages differ. In that case the source form can be substituted by a form of another type in the target language on the basis of their functional identity.

(5) Assimilation is a device applied for translating grammatical forms constituting compound structure, and the combinability features of these forms differ in the source and target languages.

(6) Conversion is used for substituting a form of one category by a form of another category, and is conditioned by the combinability rules difference in the source and target languages.

(7) Antonyms employment is used for eliminating a conflict between lexical and grammatical combinability of language units in the source and target languages.

Thus it is obvious that the search for equivalence should be carried out starting with the establishment of semantic equivalence of patterns notwithstanding their structural dissimilarity. Pattern-matching approach for the English – Russian transfer was assumed, and the segmentation of structures of the source language was performed on the basis of functional transfer fields which were established via contrastive study of the two languages.

2. Cognitive and Functional Aspects of Transfer Modelling

The machine translation technique employed presupposes three stages: analysis, transfer and generation. The stage of analysis results in parse representing the structure of the input sentences. Transfer is a bridge between the parse structure of the source language and the input to the generation procedure for the target language. At this stage the transformation is performed of one parse tree (applicable for the source language presentation) into another tree (presenting the target language). Thus syntactic transformations imply the mapping of one tree structure to another.

It is very important that *a parse for MT differs from parses required for other purposes*. Thus the grammar formalisms developed for a unilingual situation (phrase stucture rules systems for the English language) [Grover et al., 1993] would give an untransferable parse in many crucial situations. For example, just one English phrase structure rule for simple sentence would suffice for grammar parse without translation, but for the English – Russian transfer a multiple structure of possible parses is required depending on the specific finite verbal form constituting the sentence. And to overcome this, an accurate scheme for all the particular verbal form cases should be designed.

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

And 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. However such transformation proved to be of considerable danger to the whole sentence structure and might cause an unpredictable generation result. Hence, for many cases a more clumsy, though robust method of a passive construction generation was accepted: the one similar to the English "*be* + Past Participle":

V(aux_ppt) → V(aux) PPt

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). *Only by creating the centaur concepts.. 'constituency-dependency', 'linearity-nonlinearity', 'form-function', etc. can we get a reasonably clear picture of linguistic reality* [Shaumyan, 1987].

The starting idea for the language structures segmentation strategy was the notion of functional semantic fields. The system of grammar units, classes and categories with generalized content supplementary to the content of lexical units, together with the rules of their functioning, is a system which in the end serves for transmission of generalized categories and structures of mental content which lie the foundation of utterance sense, and constitute the basis of language grammar formation [Bondarko, 2001].

As it was exhibited in [Kibrik, 2001] language coding technique is to a great extent determined by the deep semantic structure, and of considerable advantage is such a presentation method which takes for the starting point the semantic level, and particular semantic units are confronted with the coding devices expressing them. The approach of functional semantics concords in many aspects with the categorial grammar. The system of sentence members (functional roles) is being modified, but its essence is preserved in the new facts qualification via the traditional categories [Zolotova, 2001].

The transferability of phrase structures is conditioned by the choice of language units in the source and target languages belonging to the same Cognitive Transfer Spaces (CTS), notwithstanding the difference or coincidence of their traditional categorial values. A set of basic CTS was singled out and language patterns employed for conveying the functional meanings of interest were examined.

- Primary Predication CTS (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 CTS bearing the features of verbal modifiers for the Primary Predication CTS. 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 CTS: language structures performing the nominative functions (including the sentential units) comprise this field.
- Modality and Mood CTS: 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 CTS: included here are lexical – syntactic means employed for concatenation of similar syntactic groups and subordination of syntactic structures.
- Attributiveness CTS: 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 CTS: this field comprises language means for presenting entities in terms of parameters and values, measures, numerical information.
- Partition CTS: included in this field are language units and phrase structures conveying partition and quantification (e.g. *some of, part of*, *each of*, etc.).
- Orientation CTS: this field comprises language means for rendering the meaning of space orientation (both static, and dynamic).
- Determination CTS: 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 CTS: 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 CTS: lexical – syntactic structures conveying negation (e.g. *nowhere to be seen*, etc.).
- Reflexivity CTS: this field is of specific character since the transfer of reflexivity meaning goes across lexical - syntactic – morphological levels.
- Emphasis – Interrogation CTS: language means comprising this field are grouped together since they employ grammar inversion in English.
- Dispersion CTS: individual language structures specific for a given language are included here; these are presented as phrasal templates which include constant and variable elements.

The set of functional meanings together with their categorial embodiments serve 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. To implement the feature-valued inheritance sometimes broader contexts are taken.

3. Statistical Approach to Machine Translation

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 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 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 provides 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, *N*-grams of variable length in particular, introduction of semantic information into *N*-grams. 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.

4. 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*:

$$\arg\max_{A} P(A \mid S,T) = \arg\max_{A} P(A,S,T) \qquad (1)$$

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 $(B_{1,\dots,}B_K)$. The probability of each sequence is independent from the probabilities of other sequences, and it depends on the sentences in the given sequence only [*Gale, Church, 1993*]. Then

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

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 languages 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 [Chen, 1993] 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

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

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

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

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 ad 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:

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

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. For a particular alignment *m* probabilities of translations are multiplied, and the individual translations are independent one from another.

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 [Masahiko, Yamazaki, 1996] 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 methods of parallel texts alignment for creating statistical translation models were, as a rule, developed on the basis of word matches: each word in the chain of a source text had to be matched with the corresponding word in the chain of the target text (in the language of translation) and vice versa. However, quite often it is difficult to establish which words of the target and source chains correspond to each other. Special problems arise when attempting to align the words inside idioms, in case of translational paraphrases, in free translation and when the auxiliary words are omitted. The alignment of two word chains can be quite sophisticated. It is necessary to take into account various transpositions of words, omissions, insertions, and the alignments between different language levels: when a word in the source text corresponds to a phrase in the target text, and the opposite situation. The most general definition of the word-based alignment is given in [Och, Ney, 2000]. The phrase-based translation moodel, or the alignment template model [Och, Ney, 2004] 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 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.

5. 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) [Kozerenko, 2003], [Kozerenko,2008] 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* [Kozerenko, 2008]. 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 :

$$T \to \langle \rho, \alpha, \sim \rangle, \qquad (6)$$

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 $\sim$ 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_ρ 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.

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 [Och, Ney, 2004], 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.

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.

Cognitive Transfer Grammar G_{CT} is a set

.

$$G_{CT} = \{T_{L_1}, T_{L_2}, N_{L_1}, N_{L_2}, P_{CA}, P_{CT}, S_{L_1}, S_{L_2}, \mathrm{M}, \mathrm{D}\} \qquad (7)$$

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 nonterminal 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 used as linguistic filters in 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 and the mechanisms of disambiguation basing on the 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.

CONCLUSION

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. One of the latest developments is connected with implementing the natural language web sevice for the multilingual search and analysis of financial information.

The Cognitive Transfer approach provides a sound and extensible platform for simulation of cross-lingual syntactic-semantic transfer and can be applied to a greater number of languages (especially with similar categorial feature-value structures). However, the problems of discontinuity, reference resolution and ambiguity , though partially treated, still remain. Further research is connected with introducing special feature-value augmentations to the existing presentations for tracing the discontinuous structures, specifying the semantic values of particular head features and verbal subcategorization frames, and numerous phrasal units adjustment.

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.

We have evidence that by focusing on the *cognitive transfer principles* we will be able to build natural language translation systems which are more accurate, efficient, and scalable than those which currently exist. It is the goal of the current development to advance this method by means of the language engineering environment developed in the course of the current project.

The approach taken would be important in further development of educational programs for computer science and computational linguistics courses. Educational relevance of the methods proposed lies in deeper understanding of uniform cognitive mechanisms employed in particular language embodiments of semantic structures.

The work was supported by the Russian Foundation for Basic Research, Grant 11-06-00476-a "Cognitive Linguistic Representations and Disambiguation of Language Structures for Intelligent Knowledge Management and Machine Translation Systems".

BIBLIOGRAPHY LIST

[Nirenburg, et al.,1992] Nirenburg, S., Carbonell, J., Tomita, M., and Goodman, K. Machine Translation: A Knowledge-based Approach. Morgan Kaufmann. 1992.

[Visson, 1989] Visson, L. Syntactical Problems for the Russian-English Interpreter. No Uncertain Terms, FBIS, vol. 4, N 2, 1989, 2-8.

[Visson, 1991] Visson, L. From Russian Into English: An Introduction to Simultaneous Interpretation. Ann Arbor, Michigan: Ardis, 1991.

[Grover et al., 1993] Grover, C., Carroll, J. and Briscoe, T. The Alvey Natural Language Tools Grammar (4-th Release). Technical Report, 1993, Computer Laboratory, University of Cambridge, 1993.

[Shaumyan, 1987] Shaumyan, S. A Semiotic Theory of Language. Indiana University Press, 1987.

[Bondarko, 2001] Bondarko A.V. Printsipy Funktsional'noi Grammatiki I Voprosy Aspektologhii. Moskwa, URSS, 2001 /Functional Grammar Principles and Aspectology Questions. Moscow, URSS, 2001 (In Russian).

[Kibrik, 2001] Kibrik A.E. Ocherki po Obstchim I Prikladnym Voprosam Yazykoznaniya. Moskwa, URSS, 2002. /Studies in General and Applied Linguistics Issues. Second Edition. Moscow, URSS, 2001 (In Russian).

[Zolotova, 2001] Zolotova G.A. Kommunikativnye Aspekty Russkogo Sintaksisa. Moskwa, URSS, 2001/ Communicative Principles of the Russian Syntax. Moscow, URSS, 2001 (In Russian).

[Gale, Church, 1993] Gale W. A., Church K. W. A program for aligning sentences in bilingual corpora // Computational Linguistics, 1993. Vol. 19. P. 75–102.

[Chen, 1993] 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.

[Masahiko, Yamazaki, 1996] Masahiko H., Yamazaki T. High-performance bilingual text alignment using statistical and dictionary information // ACL 34, 1996. P. 131–138.

[Och, Ney, 2000] Och F. J., Ney H. A comparison of alignment models for statistical machine translation // COLING'00: The 18th International Conference on Computational Linguistics. Saarbrucken, Germany, 2000. P. 1086–1090.

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

[Kozerenko, 2003] 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.

[Kozerenko, 2008] 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.

КОГНИТИВНО-ЛИНГВИСТИЧЕСКИЕ ПРЕДСТАВЛЕНИЯ ЯЗЫКОВЫХ СТРУКТУР В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ПЕРЕВОДА

Козеренко Е.Б.

Учреждение Российской академии наук
Институт проблем информатики РАН,
Москва, Российская Федерация

kozerenko@mail.ru

В работе рассматриваются вопросы создания представлений семантики синтаксических структур для русского и английского языков в системах машинного перевода и обработки знаний. Рассматривается проблема сегментации предложений для установления фразовых структур, переводимых методом трансфера. Применяется подход на основе обобщенных когнитивных сущностей, которые проявляются, как в системах грамматических категорий ряда европейских языков, так и в функциональных ролях языковых единиц в предложении.

Разработан и реализован декларативный модуль синтаксического анализа и синтеза системы машинного перевода “Cognitive Translator”; данный подход также использовался при создании лингвистических процессоров интеллектуальных систем обработки знаний.

Keywords: машинный перевод, фразовые структуры, синтаксис, семантика, трансфер.

Введение

Данная работа посвящена актуальным проблемам создания семантико-синтаксических представлений для систем машинного перевода и извлечения знаний из естественно-языковых текстов. Целью наших исследований является построение целостной лингвистической модели на основе синергетического подхода, использующего лингвистические знания, статистические методы и механизмы машинного обучения для извлечения новых грамматических правил из текстовых корпусов и разрешения неоднозначности. Для формализации лингвистических знаний используется когнитивная трансферная грамматика (КГТ), являющаяся семантически мотивированным вариантом унификационно-порождающей грамматики. Для подготовки обучающих компонентов систем и получения статистических данных о языковых структурах создается многоязычный лингвистический ресурс, представляющий собой банк синтаксических деревьев (Treebank) и корпус семантически выровненных параллельных текстов на русском, английском и ряде других европейских языков.

Основное содержание

Современный период развития исследований и разработок в области машинного перевода и систем извлечения знаний из текстов характеризуется интенсивным процессом «гибридизации» подходов и моделей. Потребность в этом носит объективный характер. Значительные вычислительные ресурсы современных систем позволяют накапливать и использовать ранее переведенные текстовые фрагменты, обеспечивать машинный перевод, основанный на прецедентах эффективно поддерживать компоненту «переводческой памяти».

Для машинного перевода наиболее сложной проблемой является реализация языковых трансформаций, которые необходимо производить при переводе с одного языка на другой. Текущий этап развития систем машинного перевода характеризуется исследованиями в области когнитивной семантики, вероятностных языковых моделей и разработкой семантико-синтаксических представлений, учитывающих многозначность и неоднозначность синтаксических структур.

Предлагаемый нами подход на основе когнитивной трансферной грамматики (КТГ) дает возможность компактного представления структуры составляющих предложения (грамматика фразовых структур), с одной стороны, а, с другой стороны, учитывает механизмы зависимости между узлами дерева предложения. Ядро КТГ составляют прототипические структуры исследуемых языков (в исходной модели – русского и английского), их наиболее вероятные позиции в предложении, а также статистические данные о дистрибутивных характеристиках структур (т.е. информация о контекстных условиях употребления исследуемых объектов - о структурных контекстах), схемы полного разбора предложений.

Заключение

Система когнитивной трансферной грамматики, дает возможность строить такие алгоритмические представления, которые не ведут к экспоненциальному росту правил и вычислительных затрат.

Дальнейшие исследования связаны с расширением числа типов трансформаций в англо-русском и русско-английском переводе и построением лингвистических представлений для многоязычной ситуации.

Работа выполнена при поддержке гранта РФФИ 11-06-00476-а «Когнитивно-лингвистические представления и разрешение неоднозначности языковых структур в системах интеллектуальной обработки знаний и машинного перевода».