

Semantic Approach to Language Structures Presentation for Machine Learning Algorithms Design

Elena Kozerenko

Institute for Informatics Problems of the Russian Academy of Sciences
44 corpus 2 Vavilova Str., Moscow, 119333, Russia
kozerenko@mail.ru

Abstract *The problem of establishing transferable language structures is considered. The key idea is developing a synergistic approach combining semantic grammar rules with the machine learning mechanisms of grammar rules extraction from parallel text corpora. The predesigned rules are founded on the unified cognitive structures extracted from the systems of grammar categories of the Russian and English languages and functional roles of language structures in a sentence. Machine learning methods are used to establish the weights of the meaningful language units and structures for probabilistic augmentation of the rule system for syntactic – semantic sentence analysis. The formalism employed for presentation of the English-Russian matches is a unification grammar variant.*

Keywords: *machine translation, syntax, semantics, transfer, probabilistic parsing, machine learning*

1. Introduction

The approach to natural language presentations proposed in this paper consists in synergistic employment of linguistic rule-based methods presenting the core grammar for the tasks of machine translation, and machine learning techniques for new grammar rules acquisition and disambiguation of structures.

The following basic features of the given work differentiate it from the existing rule-based and learning-based approaches:

- the emphasis on the semantics of grammar, i.e. the study of language configurations basing on the Functional Transfer Fields (FTF) [1] and their projections into particular language structures; this will result in developing a computational variant of a multilingual semantic grammar (MSG);
- MSG development is dominant to the lexical semantic studies which are conducted on the basis of the existing English-Russian computational vocabulary obtained from parallel text corpora;
- The semantic grammar is used for the establishment of regular correlates between structures (configurations) and lexical units, i.e. cross-level correspondences;
- the construction of the systemic cross-lingual presentation of phrase structures conveying the similar meanings in the languages under study and employing it as the core of the rule system for the algorithms of syntactic-semantic transfer in machine translation, multilingual knowledge management and information retrieval;
- the synthetic approach enables the system to generalize the rules and avoid the overgeneration of rules thus resulting in translation accuracy improvement.

The texts of business and scientific discourse have been studied, and specific structures are extracted and serve the source for further rule set development. Experiments are conducted with the texts of corporative documents and patents.

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.

2. Modern Tendencies in Machine Translation

A significant progress in the area of natural language processing and machine translation has been reached in the latest several years due to the application of methods based on machine learning and probabilistic models for language structures analysis. These methods, in particular, were successfully applied in speech-to-speech translation systems [2,3].

The two major factors also determine the direction of research developments in the area of natural language processing at the present stage: the emergence of big parallel texts corpora and significant computing resources which allow to accumulate the translated text fragments for further use (Translation Memory and Example-based Machine Translation) [4-7]

A collection of translations can be formed manually - in that case the translations will be authentic, but expenditures of labour are rather significant, or it can be prepared on the basis of big parallel texts, when patterns for translation are selected automatically. In the latter case the means for editing and filtration of translations are necessary, since a great number of noises and superfluous rules are inevitably generated.

Currently, there are also research projects developing formal models of translation [8-10], and implemented systems, such as LMT (logical-based machine translation system from English to German) [11] of IBM company.

The latest results of research as well as our experiments [12,13,1] testify about the fact, that the optimal precision of translation is obtained by the extension of traditional linguistic rule-based approaches with machine learning methods and the use of corpus statistics.

3. Semantic Match Establishing Principles

Studying the categorial and functional meanings of language structures we have established that important tools realizing these meanings are *ways of configuring* phrase structures, i.e. *linearization patterns*: possible linear sequences of language objects (of units and structures).

Semiotic linguistics [14] calls these ways of configuring *structural signs* and also introduces the concept of *superposition of functions*, presuming that every language object has its *primary* function, and shifts of meanings which occur in the acting language are superposition of secondary and other functions onto the primary one.

Our research is focused on revealing all possible types of structural signs which convey similar meanings, i.e. establishment of *syntactic synonymy*.

A classical example of syntactic synonymy is the means of expressing case meanings, e.g. morphological in the Russian language (through cases endings) and analytical in English - by means of prepositions and the order of words.

Hence, our problem is to reveal all types of structural signs and compose them into a uniform system of semantic syntactical representations for a language processor.

Superposition of functions is a very useful tool for construction of functional semantic representations of linguistic units and structures.

The concepts of primary and secondary functions of language signs enable us to create the new representations using traditional categories (as, for example, the Verbal _ Noun = Verb + Noun).

The similar method was applied for creating a system of rules for functional transfer in machine translation [1,15]. 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. 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 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.

Our approach is founded on the previously established group of Functional Transfer Fields (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 Cognitive Transfer Grammar [1] was developed. It comprised about 300 transferable phrase structures together with the transfer rules combined within the same pattern. Such patterns, or Cognitive Transfer Structures (CTS), served constitutional components of the declarative syntactical processor module and encoded 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

These phrases have the same meaning, that could be presented by the following semantic network: Owner ← Having → Had Thing.

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

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. In Contrastive Linguistics the notion of *tertium comparationis* is widely employed to denote this general equivalent, and the approach based on the principle “from the meaning to the form” focusing on Functional Syntax would yield the necessary basis for equivalence search. We offer an alternative to the standard generative view of syntax: syntactic phenomena are presented from a number of languages and the emphasis is laid on the major typological issues that syntactic theories must address.

What differs our approach is the attention to the *semantics of configurations*, i.e. the study of the way languages tend to *arrange* structures in order to convey certain meanings. And we focus on the *linear* patterns of the languages under study, since we assume that linearization is not a random process but it is determined by the cognitive mechanisms of speech production and the way they manifest themselves in syntactic potentials of a given language. The primary object of our contrastive language study was to establish what particular language meanings are represented in the categorial-functional systems of the English and Russian languages. Categorial values embody the syntactic potentials of language units, i.e. their predictable behavior as syntactic structures (syntaxemes). Thus we can say that Category is the potential for Function [14,-16], and multiple categorial values inflict multiple syntactic functions. However, when we analyze language in action, i.e. the utterances of written or sounding speech, the function of a particular language structure determines which of the potentials is implemented in this utterance, hence which of the possible categorial values of a polysemous syntactic structure is to be assigned to the parse result.

4. The methods of machine learning in natural language processing

The aim of machine learning is to infer automatically a model for some domain on the basis of the given data from the domain, thus a system learning syntactic rules would be supplied with a set of phrase structure rules to be used for training. Recently more attention has been paid to the construction of N-grams capturing sophisticated presentations of syntactic and semantic structures: using long-distance triggers instead of local N-grams [17,18], applying variable-length N-grams [19], including semantic information to the N-grams, e.g. semantic word associations based on the latent semantic indexing [20].

Different stochastic taggers appeared in the 1980s [21,22]. The idea shared by all stochastic taggers consists in choosing the most likely tag for a given word.

One of the most popular probabilistic taggers is the Hidden Markov Model (or HMM tagger) - for a given sentence or word sequence, HMM taggers pick the tag sequence that maximizes the following formula: $P(\text{word} | \text{tag}) * P(\text{tag} | \text{previous } n \text{ tags})$.

An approach to machine learning based on rules and stochastic tagging is known as the Transformation-Based Learning (TBL). TBL is a supervised learning technique, and employs a pre-tagged training corpus.

For probabilistic parsing stochastic grammars are applied. A Probabilistic Context-Free Grammar (PCFG), is a 5-tuple $G = (N, T, P, S, D)$, where N is a set of non-terminal symbols, T is a set of terminal symbols, P is a set of productions of the form $A \rightarrow b$, where A is a non-terminal symbol, b is a string of symbols, S is a designated start symbol, D is a function assigning probabilities to each rule in P .

Probabilistic tree substitution grammar (PTSG) : the definition is the same as PCFG but rather than a set of rules, we have a set of tree fragments of arbitrary depth whose top and interior nodes are nonterminals and whose leaf nodes are terminals or nonterminals, and the probability function assigns probabilities to these fragments. PTSGs are thus a generalization of PCFGs, and are stochastically more powerful, because one can give particular probabilities to fragments - or just whole parses - which cannot be generated as a multiplication of rule probabilities in a PCFG. Bod [23] shows that by starting with a PCFG model of depth 1 fragments and then progressively allowing in larger fragments parsing accuracy does increase significantly (this reflects the result of Johnson [24] on the utility of context from higher nodes in the tree). So this model provides another way to build probabilistic models that use more conditioning context. A profound survey of statistical natural language processing methods is given in [25]. The results of semantic equivalence establishment via stochastic apparatus and alignment experiments are given in [26-29]. A very significant result is presented in [30] Collins builds more complex models capturing some of the statistical dependencies between different dependents of a head. What is of particular interest is that *the models start bringing in a lot of traditional linguistics*.

We considered TBL, PCFG and PTSG when working out the way how to introduce stochastic learning apparatus into the system of our grammar. At present an accepted approach is the probabilistic functional tree substitution grammar (PFTSG) mechanism operating with the system of multivariant semantic configurational rules.

5. The Probabilistic Model of Multivariant Syntactic Parse

The probability values for syntactic analysis variants can be obtained either on the basis of corpus information, or from linguistic expert knowledge. In the latter case we deal with reliable information fixed in grammar systems of languages distilled by the centuries of human language practice.

The values of probabilities for every possible parse variant (i.e. the expansion of a nonterminal node) are calculated on the basis of frequencies of occurrence of each analysis variant in the existing text corpora with syntactic mark-up (treebanks). The calculation is made of the number of times (N), when some variant of expansion of a node ($\alpha \rightarrow \beta$) is used with subsequent normalization:

$$P(\alpha \rightarrow \beta | \alpha) = \frac{N(\alpha \rightarrow \beta)}{\sum_{\gamma} N(\alpha \rightarrow \gamma)} = \frac{N(\alpha \rightarrow \beta)}{N(\alpha)} \quad (0.1)$$

Consider, how probability values are used in the process of parsing. For example, the Probabilistic Context Free Grammar (PCFG) and the Probabilistic Tree Substitution Grammar (PTSG) assign probability (P) to each tree of analysis T (i.e. to every derivative) of a sentence S . This information is the key for disambiguation of syntactic structures. The probability of every possible parse tree T is determined as the product of probabilities of all rules r , being used for the expansion of every node n in the parse tree:

$$P(T, S) = \prod_{n \in T} p(r(n)) \quad (0.2)$$

The resulting probability $P(T, S)$ is a joint probability of a given parse variant and a given sentence on definition of joint probability:

$$P(T, S) = P(T)P(S | T) \quad (0.3)$$

But, since the parse tree includes all words of the given sentence, then

$$P(T, S) = P(T)P(S | T) = P(T) \quad (0.4)$$

The probability of an unambiguous sentence (i.e. a sentence, where we do not need to resolve ambiguity) is equal to the probability of a single parse tree for this sentence, i.e.

$P(T, S) = P(T)$. But the probability of an ambiguous sentence is equal to the sum of probabilities of all possible parse trees ($\tau(S)$) for the given sentence:

$$P(S) = \sum_{T \in \tau(S)} P(T, S) = \sum_{T \in \tau(S)} P(T) \quad (0.5)$$

The probability of the full parse of a sentence is calculated with the account of categorial information for each head vertex of every node. Let n be a syntactic category of some node n , and $h(n)$ is the head vertex of the node n , $m(n)$ is a mother node for the node n , hence, we will calculate the probability $p(r(n)|n, h(n))$, for this we transform the expression (1.2) in such a way, that every rule becomes conditioned by its head vertex:

$$P(T, S) = \prod_{n \in T} p(r(n) | n, h(n)) \times p(h(n) | n, h(m(n))) \quad (0.6)$$

Since the ambiguity of some syntactic structure (a node) is understood as an opportunity of realization of more than one categorial value in the head vertex of this structure, the probability of the full parse of a sentence containing ambiguous structures (i.e. nodes, subtrees) will be calculated with the account of the probabilities of the categorial values realized in the head vertices of these structures (nodes).

In our grammar system the functional values of languages structures are determined by the categorial values of head vertices. The probabilities are introduced into the rules of the unification grammar (PFTSG) as the weights, assigned to parse trees. Ambiguous and polysemous syntactic structures are modeled by the Multivariant Cognitive Transfer Structures (MCTS).

6. Language Structures Presentations in Polyglot

Our linguistic simulation efforts are aimed at capturing the cross-level synonymy of language means and present them as interlingual *semantic configurational* matches. Our current project POLYGLOT is aimed at creation of systemic presentations of functionally motivated semantic syntactic structures. The above stated methods are being employed for design and development of a research linguistic knowledge base POLYGLOT. It is a linguistic resource with semantic grouping of phrase structure patterns provided with the links to isosemic structures at all language levels for the whole set of languages included into the linguistic base. The categorial systems of a subset of natural languages (English and Russian, Italian and French) and functional roles of language units in a sentence are being explored on the basis of the core set of transferable language phrase structures.

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 POLYGLOT linguistic knowledge base comprises the following components:

- parallel texts database: the texts are segmented into the functionally relevant structures that are semantically aligned;
- multilingual functional Treebank (under development at present);
- structural parse editor (under development at present) which displays the parse and transfer schemes for indicated text segments;
- functional semantic vocabulary of structural configurations arranged on the basis of the Functional Transfer Fields principle.

Our linguistic simulation efforts are aimed at capturing the cross-level synonymy of language means inside a system of one natural language and interlingual *semantic configurational* matches. 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.

7. Conclusions

We see our principal objective in developing a novel synthetic approach where language structures of several natural languages are aligned on the basis of functional meanings conveyed by a concise system of categorial values presented in cross-lingual charts, lexical and structural disambiguation is performed by means of stochastic techniques and new structures and patterns are acquired with the help of machine learning methods. Our analysis and development experiments result in understanding the efficiency of what might be called the “exteriorization of meaning”, i.e. accumulation of relevant data concerning the functional-categorial features of possible structural contexts and/or specific lexical contexts that help to disambiguate the parsed structures of the source language and decide on what particular meaning of a language structure is realized in the given text segment. Rather than invent a sophisticated antropocentric heuristics for the rule-based disambiguation techniques via traditional linguistic presentations, we need to design a synthetic mechanism comprising the core rule set and reliable learning methods.

The rule set applicable both for the transfer procedures and for acquiring new linguistic data by corpora study should envisage the phenomena of syntactic polysemy and ambiguity of structures. The solution employed in our project is based on the Functional Transfer Fields approach grouping isofunctional language structures, and the Multivariant Cognitive Transfer Grammar (MCTG) comprising the rules which state the multiple equivalent structures in the source and target languages. The MCTG linguistic rule set is being augmented by Probabilistic Functional Tree Substitution Grammar (PFTSG) features. Since the nodes of the MCTG have functional articulation, the trees and subtrees of the possible parses also have functional character, i.e. are tagged by functional values.

Further research and development is connected with the refinement of the existing presentations, inclusion of specific lexical-based rules into the grammar system, and excessive corpora-based experiments for extending the mechanisms of multiple transfer.

References

- [1] 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, June, 23-26, 2003, Las Vegas, USA.// CSREA Press, pp. 49-55, 2003.
- [2] Kay, M., Gawron, J., and Norvig, P. Verbmobil: A Translation System for Face-to-Face Dialog. CSLI; 1992.
- [3] Frederking, R., Rudnicky, A.I., and Hogan, C. Interactive speech translation in the DIPLOMAT project. In Proceedings of the ACL-97 Spoken Language Translation Workshop, Madrid, pp. 61-66. ACL, 1997.
- [4] Sumita, E. and Iida, H. Experiments and prospects of example-based machine translation. In ACL-91, Berkeley, CA, pp. 185-192. ACL, 1991.
- [5] Brown, R.D. Example-based machine translation in the Pangloss system. In COLING-96, Copenhagen, pp. 169-174, 1996.
- [6] Nagao, M. A framework of a mechanical translation between Japanese and English by analogy principle. In Alick Elithorn and Ranan B. Banerji (eds.), Artificial and Human Intelligence, pp. 173-180. Edinburgh: North-Holland, 1984.
- [7] Sato, S. CTM: An example-based translation aid system. In COLING 14, pp. 1259-1263, 1992.
- [8] Alshawi, H., Bangalore, S., and Douglas, S. Automatic acquisition of hierarchical transduction models for machine translation. In COLING/ACL-98, Montreal, pp. 41-47. ACL. 1998.
- [9] Knight, K. and Al-Onaizan, Y. Translation with finite-state devices. In Farwell, D., Gerber, L., and Hovy, E.H. (Eds.), Machine Translation and the Information Soup. pp. 421-437. 1998.

- [10] Wu, D. and Wong, H. Machine translation with a stochastic grammatical channel. In COLING/ACL-98, Montreal, pp. 1408-1414. ACL, 1998], and implemented systems, such as LMT (logical-based machine translation system from English to German)
- [11] McCord Michael C. Design of LMT: A Prolog-Based Machine Translation System. <http://acl.ldc.upenn.edu/J/J89/J89-1003.pdf>
- [12] Habash, Nizar and Bonnie Dorr. Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. AMTA-2002. Tiburon, California, USA, 2002.
- [13] Dorr, Bonnie and Nizar Habash. Interlingua Approximation: A Generation-Heavy Approach. AMTA-2002 Interlingua Reliability Workshop. Tiburon, California, USA, 2002.
- [14] Shaumyan, S. Categorical Grammar and Semiotic Universal Grammar. In Proceedings of The International Conference on Artificial Intelligence, IC-AI'03, Las Vegas, Nevada, CSREA Press, 2003.
- [15] Kozerenko, E.B., Shaumyan, S. Discourse Projections of Semiotic Universal Grammar // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 27-30, 2005, Las Vegas, USA.// CSREA Press, pp. 3-9, 2005.
- [16] Shaumyan, S. Intrinsic Anomalies in Meaning and Sound and Their Implications for Linguistic Theory and Techniques of Linguistic Analysis. // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 27-30, 2005, Las Vegas, USA.// CSREA Press, pp. 10-17, 2005.
- [17] Rosenfeld, R. A maximum entropy approach to adaptive statistical language modeling. Computer Speech and Language, 10, 187-228, 1996.
- [18] Niesler, T.R. and Woodland, P.C. Modelling word-pair relations in a category-based language model. In IEEE ICASSP-99, pp. 795-798, IEEE, 1999.
- [19] Ney, H., Essen, U., and Kneser, R. On structuring probabilistic dependencies in stochastic language modeling. Computer Speech and Language, 8, 1-38, 1994.
- [20] Jurafsky, D. and Martin, J.H. Speech and Language Processing. Prentice Hall, 2000.
- [21] Marshall, I. Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB corpus. Computers and the Humanities, 17, 139-150, 1983.
- [22] Church, K.W. A stochastic parts program and noun phrase parser for unrestricted text. In Second Conference on Applied Natural Language Processing, pp. 136-143. ACL, 1988.
- [23] Bod, Rens. Enriching Linguistics with Statistics: Performance Models of Natural Language. PhD thesis, University of Amsterdam; Bod, Rens, and Ronald Kaplan. A probabilistic corpus-driven model for lexical-functional analysis. In ACL 36/COLING 17, pp. 145-151, 1995.
- [24] Johnson, Mark. The effect of alternative tree representations on tree bank grammars. In Proceedings of Joint Conference on New Methods in Language Processing and Computational Natural Language Learning (NeMLaP3/CoNLL98), pp.39-48, Macquarie University, 1998.
- [25] Manning, C.D., Schütze, H. Foundations of Statistical Natural Language Processing, 1999.
- [26] Melamed, I. D. Bitext Maps and Alignment via Pattern Recognition. Computational Linguistics 25(1): 107-130, 1999.
- [27] Gildea, D. Loosely tree-based alignment for machine translation. In Proceedings of the 41-st Annual Conference of the Association for Computational Linguistics (ACL-03), Sapporo, Japan, 2003.
- [28] Gildea, D. and M. Palmer. The necessity of syntactic parsing for predicate argument recognition. In Proceedings of the 40-th Annual Conference of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, 2002.
- [29] Gildea, D. Probabilistic models of verb-argument structure. In Proceedings of the 19-th International Conference on Computational Linguistics (COLING-02), pp. 308-314, Taipei, 2002.
- [30] Collins, Michael John. Three generative, lexicalised models for statistical parsing. In ACL 35/EACL 8, pp. 16-23, 1997.