Government Models for Cross-Lingual Transformations

Elena B. Kozerenko
Institute for Informatics Problems of the Russian Academy of Sciences, Moscow, Russia
kozerenko@mail.ru

The paper deals with the issues of creating presentations for computational models of cross-lingual correspondencies which take into account the transformations of language structures indispensable for the adequate translation. For this task of primary concern are the matters of government relations between language objects [1-8]. In this connection very important is the notion of "valence" (or "valency") which refers to the number of participants a verb requires to complete its meaning, e.g. *love* as in *Jane loves Charles.* Langacker uses the term "valence" in the meaning of what is traditionally described as the head-dependent relation [1,2]. Our view of the notions *government* and *subcategorization* could be summed up as follows: *subcategorization* is understood as the enumeration of the expected categorial features of the language objects co-occurring with the head element, while *government* specifies both categorial and morphosyntactic features of the dependent language objects.

In this paper the English-Russian language pair is considered, however our experiments show that major conclusions hold for the Byelorussian and Ukrainian languages that are closely related to Russian, as the syntactical transformation processes are similar in these languages. Mainly the syntactic aspect of cross-lingual transformations is considered here. We focus on the correspondecies between verbal and nominal units and finite-nonfinite verb forms in the source and target languages. The two presentation mechanisms are employed in our developments. The first mechanism is based on dependency grammar, and it is applied for the design of the multilingual knowledge extraction systems, several projects have been implemented on these principles [9]. The other mechanism takes into account both constituency and dependency relations, and it is used for parallel texts alignment and for the transfer-based machine translation system development, a hybrid machine translation system employing rules and statistics was realized [10]. The first mechanism is based on the extended semantic networks (ESN) which have the sufficient expressive power for presenting the highly embedded structures of natural language. The basic structural element of the ESN is the named *N-ary* predicate, called "fragment". ESN is the development of this type of networks in the direction of the descriptive power increase with the retention of uniformity. The ESN basis is the set of vertices (V), from which the following elementary fragments are comprised: *V0(V1,V2,...,Vk/Vk+1), where V0,V1,V2,...,Vk,Vk+1 V, k > 0.*

This fragment represents a *k*-ary relation. The fragments are assigned their roles. The vertex *V0* corresponds to the name of relation, the vertices *V1, V2,..., Vk* correspond to the objects which are linked by the relation, and the vertex *Vk+1* separated by the line (/) from the entire structure corresponds to the vertex of connection. The *Vk+1* is called a *C*-vertex, and all these elements form the extended semantic network (ESN). The whole set of language objects are given in the form of predicate-argument structures. The uniformity of language presentations is a very important factor, and in the process of analysis of natural language sentences the unification grammar is used. With this approach the words and the constructions, which perform the role of predicates in the sentence, serve as the "support" elements, and the result of the analysis of a sentence is one "extended" predicate, which corresponds to the predicate of a sentence (i.e. to the basic verb in the tensed form or to another basic predicate expression). The government models and transformation features are given in the vocabulary entries of verbs:

e.g. *shoot the ducks from the rifle –*
strelyat' utok iz ruzh'ia/ strelyat' po utkam iz ruzh'ia
shoot (V) ducks from rifle / shoot (V) at ducks from rifle

shooting the ducks from the rifle –
strel'ba po utkam iz ruzh'ia / strelyaiuschii po utkam iz ruzh'ia
shooting (N, process) at ducks from rifle / shooting (Part) at ducks from rifle
/ strelyaia po utkam iz ruzh'ia
shooting (AdvPart) at ducks from rifle.

The transformations result in the shift of the government models. Special attention in our research is given to the cases of nominalization and changes from prepositional government models to those without prepositions: *strelyat' po utkam - shoot the ducks.* It is vital to explore "the synonymy" of structures, i.e. all possible realizations of the "unit of sense" – the ESN structure which serves as the representation of meaning in the knowledge base. We focus on the detailed experimental study of language transformations in translated texts. The data is obtained from the text corpora of scientific articles, patents and business documents of our linguistic resource and other corpora available online. One of the basic transformations is the *nominalization*. Our research shows that the Russian language is about 35% more "nominative" than English. This information is introduced into the rules of transfer and alignment. Thus, for example the following translational correspondences occur regularly:

In vacuum molecules have large space in which to move (V).
V vakuume molekuly imeiut bol'shoe prostranstvo dlia dvizhenia. (Rus.-translit)
In vacuum molecules have large space for movement (N).

At present experiments are conducted with implementing the natural language web service for the multilingual search and analysis of financial information.

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 objective of our studies is the establishment of a linguistically motivated translation model which would serve as the basis for parallel texts alignment and automatic acquisition of rules capturing the subcategorization features for language units. Many individual syntactic objects are "incomplete" and require an argument to "flesh out" their syntactic and semantic requirements [3].

References

1. Langacker, R. (1987) Foundations of Cognitive grammar, Volume I. Stanford, CA: Stanford University Press.
2. Langacker, R. (1991) Foundations of Cognitive grammar, Volume II. Stanford, CA: Stanford University Press.
3. Carnie, A. Constituent Structure. Oxford Surveys in Syntax and Morphology. Oxford University Press. 2008.
4. Briscoe, E. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In Proceedings of the 5th ACL Conference on Applied Natural Language Processing, pages 356–363, Washington, DC.
5. Carroll, G. and Rooth, M. (1998). Valence induction with a head-lexicalized pcfg. In Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing, pages 36–45, Granada.
6. Kozerenko E.B. Multilingual Processors: a Unified Approach to Semantic and Syntactic Knowledge Presentation // Proceedings of the International Conference on Artificial Intelligence IC-AI'2001. H.R. Arabnia (ed.), Las Vegas, Nevada, USA, June 25-28, 2001. CSREA Press, 2001. P.1277-1282.
7. Johnson, C. R., Fillmore, C. J., Petruck, M. R. L., Baker, C. F., Ellsworth, M., Ruppenhofer, J., and Wood, E. J. (2006). FrameNet: Theoryand Practice. Word Wide Web, http://www.icsi.berkeley.edu/~framenet/book/book.html .
8. Sag, I. and Wasow, T. (1999). Syntactic Theory. A Formal Introduction. University of Chicago Press, Chicago.
9. Kuznetsov I.P., Kozerenko E.B., Kuznetsov K.I., Timonina N.O. Intelligent System for Entities Extraction (ISEE) from Natural Language Texts // Proceedings of the International Workshop on Conceptual Structures for Extracting Natural Language Semantics - Sense'09, Uta Priss, Galia Angelova (Eds.), Moscow, Russia, 2009. P. 17-25.
10. 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.