View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by DSpace at Tartu University Library

Identification of Context Markers for Russian Nouns

Anastasia Shimorina
St. Petersburg State University
St. Petersburg, Russia
shinas@yandex.ru

Maria Grachkova
St. Petersburg State University
St. Petersburg, Russia
maaag86@mail.ru

Abstract

The research project presented in this paper aims at identification of context markers for Russian nouns and their use in construction identification. The body of contexts has been extracted from the Russian National Corpus (RNC). The context processing procedure takes into account the lexical and semantic information represented in the corpus annotation. Merged meaning of words are taken into consideration. The reported results contribute to task of building a comprehensive lexicographic resource — the Index of Russian lexical constructions.[1]

1 Introduction

The importance of corpus data is now widely recognised. The corpus shows functioning of language units in their natural domain of occurrence and it serves for various linguistic tasks (e.g., (Rakhilina et al., 2006)). This research project uses the Russian National Corpus (RNC, http://www.ruscorpora.ru/) as a resource providing context markers of word meanings. Context marker of a target word is a linguistic unit occurring in one context with this word and specifying its particular meaning. RNC has a multilevel annotation, it includes lexical (lemma) tags (*lex*), morphological (grammatical) tags (*gr*), and semantic (taxonomy) tags (*sem*). These tags should be taken into account when operating with context markers. Context markers find an application in construction identification and word sense disambiguation (WSD) (e.g., (Agirre and Edmonds, 2007; Navigli, 2009; Mihalcea and Pedersen, 2009; Proceedings of the NAACL HLT Workshop… 2010; Sahlgren and Knutsson, 2009), etc.). Corpus-based WSD implies extraction and statistical processing of word collocations, which makes it possible to distinguish separate meanings of lexical items in context (e.g., (Kobricov et al., 2005; Lashevskaja and Mitrofanova, 2009; Pedersen, 2002; Schütze, 1998), etc.).

2 Linguistic data and experiments

Four Russian polysemous words were subjected to analysis: *organ* 'institution, part of the body, musical instrument, etc.', *luk* 'onion, bow', *glava* 'head, chief, cupola, chapter, etc.', and *dom* 'building, private space, family, etc.'. Sets of contexts were extracted from the RNC, the largest annotated corpus of Russian texts containing about 400 M tokens. We deal with the disambiguated portion of the RNC where morphological and semantic ambiguity is resolved. The size of context set for each noun ranges from 1000 to 3500. The texts are supplied with three core types of annotation: (1) lemmas — lexical markers (canonical, dictionary forms of inflected words); (2) grammatical markers (morphosyntactic tagsets referring to POS and other inflectional grammatical features like case, gender, tense, etc.); (3) taxonomy markers (semantic tagsets referring to lexical-semantic classes). Taxonomy markers are available for the most frequent nouns, pronouns, adjectives, verbs and adverbs and represent a rather coarse-grained cross-classification of the lexicon (e.g. 'concrete', 'human', 'animal', 'space', 'construction', 'tool', 'container', 'substance', 'movement', 'part', 'diminutive', 'causative', 'verbal noun', and other lexical-semantic classes, cf. http://www.ruscorpora.ru/en/corpora-sem.html). Each word sense is formalized with a set of taxonomy markers, cf. *dom* 'house': 'concrete' + 'construction' + 'container'. A list of contexts is made for each meaning of considered words.

[1] This work was funded by the Russian Foundation for Basic Research (grant No 10-06-00586-a).

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 344–347

Further, we extract automatically from these contexts the lexical-semantic and statistical information about words that are to the left (right) of the analyzed noun. This information is presented as a set of semantic tags. The semantic tagsets are arranged by their frequency of occurrence, then we consider only the statistically significant sets. The frequency tagsets are analyzed in terms of what lexical units are behind the semantic tagsets. These lexemes are most probably the context markers of the considered words.

A Python-based WSD and Construction Identification toolkit (Lyashevskaya et al., in press) was used in order to extract and analyze context markers. The toolkit makes it possible to carry out linguistic and statistical analysis of contexts for polysemous words in various modes. It performs (1) generation of context classes corresponding to particular meanings of a target word; and (2) generation of lists of the most frequent constructions where a particular meaning of a target word occurs.

3 Identification of context markers

Context markers were determined for each meaning of the words listed above. The markers can be of various nature, e.g. they may represent different parts of speech. Much attention was paid to lexical-semantic tags of context markers. For example, the target word *glava* 'chief' frequently co-occurs with the following lexemes forming its right context: *gosudarstvo* ('state' <r:concr t:space>), *federacija* ('federation' <r:concr t:space>), *region* ('region' <r:concr t:space pt:part pc:space>), *gorod* ('city' <r:concr t:space sc:constr>), *fond* ('fund' <r:concr t:space pt:set sc:money>). These context markers can be combined to form a group of concrete nouns identifying space and place (<r:concr t:space>). To take another example, the target word *luk* 'onion' regularly co-occurs with such nouns as *ogurec* ('cucumber' <r:concr t:fruit t:food>), *orekh* ('nut' <r:concr t:fruit t:food pt:part pc:plant>), and *kartoška* ('potato' <r:concr t:fruit t:food pt:aggr sc:fruit>). These nouns may be referred to as a group of concrete nouns denoting food. These examples show that the identification of context markers can be carried out not in terms of particular lexemes, but in terms of the lexical-semantic classes they belong to.

Context markers may differ not only in type, but also in the position they occupy with respect to a target word. Therefore, the right and left contexts of target words were examined separately. For instance, semantic tags indicating abstract nouns of perception (<r:abstr t:perc>) regularly occur in the right context of the target word *organ* ('part of a body'). This fact allows us to consider them as context markers for the word in question. But when we explored the left context of the same word in the same meaning, we found out that other lexemes often serve as its context markers: e.g., adjectives, such as *čelovečeskij* 'human', *donorskij* 'donor' (<dt:hum>), nouns *zabolevanije*, *bolezn'* 'disease' (<t:disease>), etc. The context markers mentioned above are not to be found in any occurrences of the word *organ* in other meanings. The combinations of target words and identified context markers are considered as constructions. The characteristic features of construction are stability and frequency of occurrence.

In order to prove the stability of obtained constructions we adopt a statistical approach. A lexeme under consideration and its context marker act as a bigram. Bigram search service (http://www.aot.ru/) provides the necessary information about the stability of bigrams. These statistical data show that the collocations have a high Mutual Information (MI), cf. Table 1.

Left context	MI	Right context	MI
pra-voohranitelnyj 'law-enforcement'	13.61	*gosbezopasnost* 'a state security'	11.23
ispolnitelnyj 'executive'	10.79	*pravoporyadok* 'law and order'	10.68
zakonodatelnyj 'legislative'	10.39	*samoupravlenie* 'self-government'	9.19
predstavitelnyj 'representative'	9.33	*zdravoohranenie* 'public health'	8.76

Table 1: Statistical results for the word *organ* 'institution'.

4 Problem of merged meanings

In automatic text processing, dictionary compiling, WSD procedure etc. linguists often have to deal with polysemous words with merged meanings. These meanings represent combinations of two or more independent meanings which are almost indistinguishable in certain contexts. In NLP tasks mentioned above such polysemous words which reveal both independent and merged meanings represent a special problem. It is hardly possible to provide unambiguous analysis of such words.

A few attempts were made in computational linguistics to solve the problem of merged meanings. For instance, the so-called "Shishkebab"

approach (Philpot et al., 2003) is used in formal ontology Omega (http://omega.isi.edu/). This method implies simultaneous attribution of two or more meanings combined in particular context to the same lexeme (e.g., *Library* IS_A Building&Institution&Location). About 400 patterns for merged meanings (e.g., *X* IS_A Country&Nation&Government; *X* IS_A Company&Product&Stock; etc.) were described in (Hovy, 2005).

This section presents the results of context markers identification experiments carried out for a noun *dom* ('house'). In the semantic structure of this polysemous noun besides six independent meanings ('building', 'private space', 'family', 'common space', 'institution', 'dynasty') there are five merged meanings formed by pairs ('building & private space', 'building & institution', 'private space & family') and triples ('building & private space & institution', 'building & private space & family') of independent ones. In the experiments we analyzed 3000 contexts for the considered noun, which were extracted from the RNC. Of the total number of contexts there are 842 contexts where the noun in question reveals merged meanings (cf. Table 2).

All occurrences of the target word found in RNC were analyzed with the exception of contexts for rare meanings found in less than 10 contexts (such as *dom* 'common place' or *dom* 'dynasty').

In the experiments we extracted lexical markers of the noun *dom* on the basis of the most frequent semantic annotation of the words adjacent to *dom*. The lexical markers of merged meanings were compared with the ones of independent meanings to decide whether additional statistical patterns should be introduced in further experiments. As the consequence of such comparison we managed to find out certain regularities in occurrence of context markers.

Some context markers allow to predict the occurrence of merged meaning with high precision. For example, context markers of merged meaning 'building & institution' are found in such pairs as *destkij dom* ('orphan's home'), *invalidnyj dom* ('home for disabled people'), *rodil'nyj dom* ('maternity hospital'), *dom otdyha* ('holiday center'), *dom kino* ('film theatre'), etc. The context marker which obviously indicates the meaning 'building & private space & family' can be found in such phrase as *hozjain doma* ('host'). However, there are context markers which indicate purely independent meanings. For example, noun *žitel'* ('tenant') in *žiteli doma* points out to meaning 'building'. In many cases such as *rodnoj dom* ('home'), *roditel'skij dom* ('one's parent's home') the merged meaning 'building & private space' is more frequent than independent. It should be noted that in the most cases we observe tendency of intersection between context markers for merged and independent meanings. For example, such adjectives as *derevjannyj* ('wooden'), *kirpičnyj* ('made of brick'), *novyj* ('new'), *sosednij* ('neighbouring'), etc. may indicate both independent meaning 'building' and merged meaning 'building & institution'.

Word meanings	Semantic annotation	Number of contexts in RNC
dom		3,000 (total)
dom 'building'	<r:concr t:constr top:contain>	1,694
dom 'private space'	<r:concr t:space>	95
dom 'family'	<r:concr t:group pt:set sc:hum>	72
dom 'common space'	<r:concr t:space der:shift der:metaph>	4
dom 'institution'	<r:concr t:org>	292
dom 'dynasty'	<r:concr pt:set sc:hum>	1
dom (merged meanings)		842
dom 'building & private space'	<r:concr t:constr top:contain | r:concr t:space>	501
dom 'building & institution'	<r:concr t:constr top:contain | r:concr t:org>	250
dom 'private space & family'	r:concr t:space | r:concr t:group pt:set sc:hum	10
dom 'building & private space & institution'	<r:concr t:constr top:contain | r:concr t:space | r:concr t:org>	36
dom 'building & private space & family'	<r:concr t:constr top:contain | r:concr t:space | r:concr t:group pt:set sc:hum>	45

Table 2: Russian noun *dom*: semantic annotation and frequencies of meanings (number of contexts in RNC).

5 Conclusion

A set of experiments on context markers identification were successfully carried out for contexts of polysemous Russian nouns which had been extracted from RNC. Different types of context markers were described.

The work demonstrates application of the obtained context markers in construction identification task. The results of experiments also reveal the necessity of special treatment of words with merged meanings and introduction of additional

statistical patterns corresponding to these meaning in different construction identification systems. Further work implies the application of the data as filters for context preprocessing and for statistical WSD.

References

Agirre E., Edmonds Ph. (eds.). 2007. *Word Sense Disambiguation: Algorithms and Applications*. Text, Speech and Language Technology, vol. 33. Springer-Verlag, Berlin, Heidelberg, New York.

Hovy E. 2005. *Ontologies (Series of Lectures)*. Vilem Mathesius Lecture Series. Prague.

Kobricov B., Lashevskaja O., and Shemanajeva O. 2005. *Sn'atije leksiko-semanticheskoj omonimii v novostnyh i gazteno-zhurnal'nyh tekstah: poverhnostnyje fil'try i statisticheskaja ocenka.* Internet–matematika 2005: Avtomatičeskaja obrabotka webdannyh. Moscow. pp. 38–57.

Lashevskaja O., Mitrofanova O. 2009. *Disambiguation of Taxonomy Markers in Context: Russian Nouns*. Jokinen, K., Bick, E. (eds.) NODALIDA 2009. NEALT Proceedings Series, volume 4, pp. 111–117.

Lyashevskaya O., Mitrofanova O., Grachkova M., Romanov S., Shimorina A., and Shurygina A. *Automatic Word Sense Disambiguation and Construction Identification Based on Corpus Multilevel Annotation.* [in press].

Mihalcea R., Pedersen T. 2009. *Word Sense Disambiguation Tutorial.* URL: http://www.d.umn.edu/~tpederse/WSDTutorial.html

Navigli R. 2009. *Word Sense Disambiguation: a Survey.* ACM Computing Surveys, 41(2), pp. 1–69.

Pedersen T. 2002. *A Baseline Methodology for Word Sense Disambiguation.* Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, February 17–23, 2002, Mexico City. pp. 126–135.

Philpot A., Fleischman M., Hovy E. 2003. *Semi-Automatic Construction of a General Purpose Ontology.* Proceedings of the International Lisp Conference. New York, NY.

Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics, pp. 25–31. Los Angeles, CA. 2010.

Rahilina E., Kobricov B., Kustova G., Lashevskaja O., and Shemanajeva O. 2006. *Mnogoznachnost' kak prikladnaja problema: leksiko-semanticheskaja razmetka v Nacional'nom korpuse russkogo jazyka.* Kompjuternaja lingvistika i intellektual'nyje tehnologii: Trudy mezhdunarodnoj konferencii Dialog 2006. Moscow. pp. 445–450.

Sahlgren M., Knutsson O. 2009. *Workshop on Extracting and Using Constructions in NLP.* NODALIDA 2009. SICS Technical Report T2009:10.

Schütze H. 1998. *Automatic Word Sense Discrimination*. Computational Linguistics, 24(1), pp. 97–123.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955