

# A Case Study of Knowledge-Rich Context Extraction in Russian

**Anne-Kathrin Schumann**

Universität Wien / Vienna, Austria

Tilde / Rīga, Latvia

anne.schumann@tilde.lv

## Abstract

This paper presents ongoing PhD thesis work dealing with the pattern-based extraction of knowledge-rich contexts (KRCs) from two Russian corpora. In recent years, the extraction of KRCs has received much attention. However, there seems to be no consolidated methodology for the automated extraction of KRCs in an application scenario and many languages lack thorough study. In this context, Russian seems to be an ideal test bed for the evaluation and improvement of KRC extraction methods, as it is typologically different from the languages studied so far and yields difficulties common to many other Slavic languages, including scarcity of processing tools and large corpora. This paper outlines the results obtained on the two corpora with shallow processing methods and discusses future improvements to the method.

## 1 Introduction and Related Work

Research in the field of knowledge-rich context extraction is related to the development of terminological knowledge bases (Meyer et al., 1992) that stress the significance of semantic relations between concepts from corpora. To this endeavour, KRCs are highly relevant, as they are naturally occurring phrases containing definitional knowledge, e. g. knowledge about semantic relations holding between concepts.

Pearson (1998) gives a first linguistic description of knowledge-rich contexts in English and subsequent studies reveal structural similarities to Pearson's results in other languages, such as Catalan (Feliu and Cabré, 2002), Spanish (Sierra et al., 2008), and French (Malaisé et al., 2005). Approaches to KRC extraction include pattern-based work (Condamines, 2002, Aussenac-Gilles and Jacques, 2006, Sierra et al., 2008, Auger and

Barrière, 2008) and bootstrapping of semantic relation markers (Condamines and Rebeyrolle, 2001) similar to methods developed for information extraction (Xu, 2007, Blohm, 2010).

Building on the standard definition of KRCs proposed by Meyer (2001), but also on related remarks in Xu (2007) and Walter and Pinkal (2006), we define knowledge-rich contexts as *naturally occurring utterances that explicitly describe attributes of domain-specific concepts or semantic relations holding between them at a certain point in time, in a manner that is likely to help the reader of the context to understand the concept in question*. KRCs can be clearly distinguished from regular definitions as well as other context types (e. g. usage examples, collocations) used in terminography.

There is no generally accepted inventory of semantic relations and the question which relations are crucial to the understanding of specific concepts is yet unanswered (Kremer and Abel, 2010). Nevertheless, it seems reasonable to define a number of target relations for the description and validation of context candidates. Building on several relation typologies that have been developed in terminology and computational linguistics (see Feliu and Cabré, 2002, ISO, 2009, Girju et al., 2007, Séaghdha and Copestake, 2009), we postulate the following target relations:

- Hyperonymy / Hyponymy
- Meronymy / Holonymy
- Process
- Causality
- Origin
- Reference
- Function

For the extraction of semantic relations in the form of KRCs, we expect to find both domain-independent and domain-specific linguistic patterns.

## 2 Extraction Experiments

### 2.1 Corpora

A series of extraction experiments was carried out to test the usefulness of pattern-based methods in Russian. Two Russian internet corpora were collected using the “Babouk” crawler (de Groc, 2011). The first corpus comprises roughly 350 000 words and covers texts from the automotive domain. Moreover, it seemed important to us to describe the behaviour of knowledge patterns in a larger and, potentially, dirtier corpus across multiple domains. Therefore, the second corpus has around 1 000 000 words and contains texts on various topics such as cars, energy supply, recycling and IT. TreeTagger (Schmid, 1994) was used for POS annotation.

### 2.2 Regular expressions

As most approaches in the field, our work is based on linguistic patterns that are detected by means of regular expressions. “Knowledge patterns” (Barrière, 2004) were defined manually by studying corpus occurrences of KRCs. The definition of knowledge patterns concentrated on linguistic predicates, as they can be understood as verbalising semantic deep-structure predicates (Sierra et al., 2008), however, non-predicative pattern candidates were also included.

Regular expressions are a powerful shallow tool for dealing with the morphological wealth of synthetic languages. However, knowledge patterns have to be combined with term representations in order to retrieve valid contexts. We experimented both with a set of specific query terms and morpho-syntactic term formation patterns that were studied in the TTC project<sup>1</sup>. While term formation patterns alone are insufficient for other tasks, such as term extraction, we believe that they constrain the sentences retrieved by the knowledge pattern in a useful way and vice versa.

### 2.3 Experimental Setup and Results

Extraction experiments included two settings. In the first setting, terms were combined with knowledge patterns in the form of regular

expressions. In the second setting, terms were replaced by morpho-syntactic term formation patterns. In both experiments, stop sentences were filtered out by means of a Perl script. The pattern matching algorithm was also implemented in Perl.

For the term setting, query terms were selected manually based on a domain-relevance judgement. Query terms include high and low frequency terms and multi-word units. Table 1 presents examples for patterns used both in the syntactic and the term setting. The knowledge pattern keyword here is *sostoit* (to consist of):

<pre>\w+ N \w+ N \w+ N ,? состо.[тц].{1,3} из</pre>
<ul style="list-style-type: none"><li>• matches all combinations of three nouns with one of the predicates: {sostoit, sostoât, sostoât’, sostoâšij} iz, including inflected forms of the participle.</li></ul>
<pre>[Тт]опливн.{1,3} шланг.{1,3}.,? состо.[тц].{1,3} из</pre>
<ul style="list-style-type: none"><li>• matches all forms of the term <i>toplivnyj šlang</i> (fuel pipe) with one of the mentioned predicates</li></ul>
<p><u>Legend:</u> sostoit iz – it consists of sostoât iz – they consist of sostoât’ iz – to consist of sostoâšij iz – the consisting of ... (participle)</p>

Table 1: Examples of regular expressions used in the syntactic and the term setting.

It can be assumed that the term setting produces more precise results, but lower recall than the syntactic setting. For a quantitative comparison of the two settings, we ran both settings on both corpora. In the term setting, 5214 regular expressions (158 query terms combined with 33 knowledge patterns) extracted 101 sentences on the car corpus, whereas 4606 regular expressions extracted 115 sentences from the multi-domain corpus. In the syntax setting, 7 term formation patterns were used to create 343 regular expressions that extracted 677 sentences from the car corpus, whereas 371 regular expressions extracted 2044 sentences from the multi-domain corpus. The syntactic setting, therefore, seems more promising in terms of recall and is

<sup>1</sup> www.ttc-project.eu. The project is funded under the European Community’s FP7/2007-2013, grant agreement n° 248005.

computationally cheaper. For the evaluation of our patterns we concentrated on the syntactic setting. Another reason was that the term setting is biased by term frequency and does not allow for the general evaluation of a pattern.

For evaluation purposes, we manually annotated 335 KRCs in the car corpus and 422 KRCs in the multi-domain corpus. We ran the syntactic setting over the samples of both corpora that contained our target KRCs. Extraction results were compared to our annotation and precision was calculated for each pattern separately. Table 2 presents results for the patterns that worked reasonably well on at least one of our corpora. In each line of the table, the first number gives the result obtained on the car corpus, whereas the second number indicates the value for the multi-domain corpus. *pattern*<sup>1</sup> indicates that the term representation in the regular expression is positioned before the knowledge pattern. *pattern*<sup>2</sup> indicates the post position of the term representation. If no number is given, *pattern*<sup>1</sup> is the default value.

Pattern keyword	Relation	Valid KRC	Invalid sentences	Precision
Sostoit <sup>1</sup>	Meronymy	27	10	<b>0,73</b>
		41	42	<b>0,49</b>
Obrazuet	Meronymy	1	0	<b>1,00</b>
		3	4	<b>0,43</b>
Vklûčâet v sebâ	Meronymy	2	1	<b>0,67</b>
		4	4	<b>0,50</b>
Osnašen	Meronymy	3	1	<b>0,75</b>
		3	5	<b>0,38</b>
Sostoit <sup>2</sup>	Meronymy	31	14	<b>0,69</b>
		47	77	<b>0,38</b>
Osnašaetsâ	Meronymy	3	2	<b>0,60</b>
		3	6	<b>0,33</b>
Ustanavlivaût	Meronymy	4	4	<b>0,50</b>
		3	27	<b>0,10</b>
Predstavlâet soboj	Reference	14	2	<b>0,88</b>
		16	0	<b>1,00</b>
Nazyvaetsâ	Reference	2	1	<b>0,67</b>
		8	3	<b>0,73</b>
Poni maetsâ	Reference	No occ. 2	No occ. 0	No occ. <b>1,00</b>
Različaût	Hyponymy	2 10	0 5	<b>1,00</b> <b>0,67</b>

Podrazdelâût	Hyponymy	No occ. 6	No occ. 3	No occ. <b>0,67</b>
Razdelâût	Hyponymy	No occ. 4	No occ. 0	No occ. <b>1,00</b>
Klassificiruet sâ	Hyponymy	No occ. 2	No occ. 0	No occ. <b>1,00</b>
Služit	Function	14 21	8 10	<b>0,64</b> <b>0,68</b>
Prednaznačen <sup>1</sup>	Function	9 17	5 14	<b>0,64</b> <b>0,55</b>
		No occ. 23	No occ. 7	No occ. <b>0,77</b>
Osušetvlâetsâ	Function	16 24	9 38	<b>0,64</b> <b>0,39</b>
Privoditsâ	Process	3 12	1 2	<b>0,75</b> <b>0,86</b>
		4 8	1 7	<b>0,80</b> <b>0,53</b>
Vozdejstvuet	Process	4 8	1 7	<b>0,80</b> <b>0,53</b>
				<b>0,73</b> <b>0,62</b>

Table 2: Precision per pattern obtained on two Russian corpora in the syntactic setting. Recall on the car corpus amounted to 0,50. Recall on the multi-domain corpus was 0,60.

### 3 Discussion and Future Work

The results presented in this paper show that shallow processing techniques work reasonably well for Russian. There is reason to assume that precision can be improved by elaborating our extraction patterns and processing tools. Recall, however, is more difficult to deal with. The reasons for missed hits consist in variation of the syntactic structure of the sentences (e.g. by means of the inclusion of subconstituents) on the one hand and complex morphological variation patterns of our extraction keywords on the other hand. If all the wealth of KRCs from corpora of highly inflectional languages is to be exploited, this matter has to be taken care of systematically. Moreover, recall is low due to the absence of extraction patterns from our manually created

list. In order to tackle this and also the first issue, automated pattern acquisition techniques need to be studied. A dynamic understanding of linguistic patterns seems reasonable also with respect to the corpus dependence of patterns pointed out by Condamines (2002).

## Acknowledgements

The research described in this paper was funded under the CLARA project (FP7/2007-2013), grant agreement n° 238405.

## References

- Alain Auger and Caroline Barrière. 2008. Pattern-based approaches to semantic relation extraction. A state-of-the-art. *Terminology* 14 (1): 1-19.
- Anne Condamines. 2002. Corpus analysis and conceptual relation patterns. *Terminology* 8 (1): 141-162.
- Anne Condamines, Josette Rebeyrolle. 2001. Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB). Didier Bourigault, Christian Jacquemin, Marie-Claude L’Homme (eds.): *Recent Advances in Computational Terminology*. (Natural Language Processing 2). John Benjamins. Amsterdam, Philadelphia: 127-148.
- Caroline Barrière. 2004. Knowledge-rich Contexts Discovery. 17th Canadian Conference on Artificial Intelligence. Mai 2004, London, Ontario, Kanada, 1-12, <http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rt doc&an=5765481&lang=en>.
- Clément de Groc. 2011. Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology, August 2011, Lyon, France.
- Diarmuid Ó Séaghdha and Anne Copestake. 2009. Using lexical and relational similarity to classify semantic relations. 12th Conference of the European Chapter of the ACL, March 30-April 3, 2009, Athens, Greece: 621-629.
- Fei-Yu Xu. 2007. *Bootstrapping Relation Extraction from Semantic Seeds*. PhD Thesis. Saarland University Saarbrücken, Uszkoreit.
- Gerardo Sierra, Rodrigo Alarcón, César Aguilar and Carme Bach. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology* 14 (1): 74-98.
- Gerhard Kremer and Andrea Abel (2010): Semantic Relations in Cognitive eLexicography. XIV Euralex International Congress, Ljouwert, Netherlands, 2010: 380-388.
- Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. International Conference on New Methods in Language Processing 1994, Manchester, England: 44-49.
- Ingrid Meyer. 2001. Extracting knowledge-rich contexts for terminography. Bourigault, Jacquemin, L’Homme (eds.): 279-302.
- Ingrid Meyer, Douglas Skuce, Lynne Bowker and Karen Eck. 1992. Towards a New Generation of Terminological Resources: An Experiment in Building a Terminological Knowledge Base. COLING 1992, August 23-28, 1992, Nantes, France: 956-960.
- International Organization for Standardization. 2009. International Standard ISO 12620: 2009 – *Terminology and Other Language and Content Resources – Specification of Data Categories and Management of a Data Category Registry for Language Resources*. ISO. Geneva.
- Jennifer Pearson, (1998). *Terms in Context*. (Studies in Corpus Linguistics 1). Amsterdam/Philadelphia: John Benjamins.
- Judit Felíu and Maria Teresa Cabré (2002). Conceptual relations in specialized texts: new typology and an extraction system proposal. TKE 2002, August 28-30, 2002, Nancy, France : 45-4.
- Nathalie Aussenac-Gilles and Marie-Paule Jacques. 2006. Designing and Evaluating Patterns for Ontology Enrichment from Text. Lecture Notes in Artificial Intelligence 4248: 158-165.
- Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. June 23-24, 2007, Prague, Czech Republic.
- Sebastian Blohm. 2010. Large-Scale Pattern-Based Information Extraction from the World Wide Web. PhD Thesis. Technical University Karlsruhe, Studer.
- Stefan Walter and Manfred Pinkal. 2006. Automatic Extraction of Definitions from German Court Decisions. Workshop on Information Extraction beyond the Document, Sydney, Australia, July 2006: 20-28.
- Véronique Malaisé, Pierre Zweigenbaum and Bruno Bachimont. 2005. Mining defining contexts to help structuring differential ontologies. In: *Terminology*, 11 (1): 21-53.