Extraction of Knowledge-Rich Contexts in Russian – A Study in the Automotive Domain

Anne-Kathrin Schumann
Universität Wien / SIA Tilde
Vienna, Austria / Riga, Latvia
anne.schumann@tilde.lv

Abstract

This paper presents ongoing research aiming at the automated extraction of knowledge-rich contexts (KRCs) from a Russian language corpus. The notion of KRCs was introduced by Meyer (2001) and refers to a term's co-text (Sebeok, 1986) as a reservoir of potentially important information about a concept. From a terminological point of view, it seems that KRCs contain exactly the kind of information that should be included into a terminology database. Accordingly, the question how KRCs can be automatically acquired has been widely studied in recent years. However, many languages including Russian still lack thorough study. This paper presents preliminary experimental results obtained on a specialized corpus in the automotive domain.

1 Shifting paradigms in terminology: dealing with contexts

Terminology studies today are marked by a notable shift of paradigms. The increasing use of corpora has not left the discipline untouched and triggered research mainly in the field of terminology extraction (cf. Ahmad and Rogers, 2001). Work on context extraction is a rather recent development, but the idea that a term's co-text yields not only linguistic, but also semantic information and corpora can be used for conceptual analysis is now widely accepted. Accordingly, Dubuc and Lauriston (1997) describe defining and explanatory contexts for terminology, Pearson (1998) provides a detailed study of defining contexts in English and ISO 12620: 2009 (ISO 2009) describes context types similar to those put forward by Dubuc and Lauriston. However, actual implementations of KRC extraction are still rare and many major languages have not been studied yet. Moreover, important theoretical and methodological issues remain unresolved. These include questions concerning the epistemological status of automatically extracted information and the notion of "concept" in a corpus-based setting. Aussenac-Gilles *et al*. (2000), for example, define the concept as a "normalized meaning", i. e. the result of corpus-based processes rather than stable, text-independent notions. It still needs to be shown how these developments relate to practical terminology work.

Our research aims at tackling these issues by giving an evaluation of corpus-based techniques in context extraction as well as by contributing to their further development. In the following section, we outline main directions of research in KRC extraction. Section 3 presents preliminary experimental results. Section 4 summarizes the results obtained and outlines further work.

2 Related work

In KRCs, knowledge about a concept's attributes or the relations it forms with other concepts is made explicit by means of cue words or other linguistic patterns (Meyer, 2001, Jacquemin and Bourigault, 2003). These can be referred to as Knowledge Patterns or KPs (Barrière, 2004). The following approaches to context extraction can be differentiated:

- pattern-based approaches: The use of linguistic patterns for context extraction was suggested by Hearst (1992) and consists in defining lexico-syntactic patterns that indicate a semantic relationship. Studies in this tradition are Pearson (1998), Meyer (2001), Barrière (2004), Malaisé *et al*. (2005), Aussenac-Gilles and Jacques (2006), Sierra *et al*. (2008), and others.

- bootstrapping of semantic relations: This method starts from pre-defined patterns or seed relations and derives new relation instances for an iterative process of pattern generalisation. Examples are, again, Hearst (1992), Brin (1998), Condamines and Rebeyrolle (2001), Agichtein and Gravano (2000), Alfonseca *et al*. (2006), Xu (2007), and Auger and Barrière (2008).

Various approaches that combine linguistic information with machine learning have been developed (Maedche, Staab, 2000; Buitelaar *et al*. 2004). A particularly interesting approach is presented by Mustafaraj *et al*. (2006) who map semantic information on frame-semantic representations and use machine learning for automated role annotation.

3 KRCs in Russian: method outline

Although frame-semantic (Fillmore, 1985) methods seem to be linguistically sounder than patterns which give the impression of being *ad hoc* constructions, they exhibit serious drawbacks. Frame representations are not readily available for many languages. In multilingual and multidisciplinary terminology, therefore, the use of robust patterns that can be easily adapted to new domains and languages seems to be more feasible.

In our study, a list of tentative patterns was created by analysing relevant contexts in specialized texts. A specialized corpus was built using the BABOUK crawler (TTC, 2010). The Russian automotive corpus spans roughly 350 000 words in plain text. On this corpus, a series of extraction experiments was carried out. A Perl script was used to extract sentences containing previously defined patterns. Pattern occurrences were counted and relevant occurrences measured against overall occurrences. This method was proposed by Barrière (2004) similarly to traditional precision metrics.

In a first experimental cycle, extraction was based on simple keyword search. Precision for all KPs was between 0,40 and 0,60. For a second cycle, 159 target terms were selected from the corpus and combined with refined patterns. Regular expressions were used for extraction in order to retrieve inflected forms. Consequently, the detected KPs should be regarded as semantic paradigms rather than lexical units. The final list of regular expressions contains 5212 items and is based on 22 KPs with the term in pre- and 11 KPs with the term in postposition. Table 1 visualizes keywords used for KP definition:

Key-word	**Transla-tion**	**Context type**	**Corpus occurr-ences**
obespeči-vaet	provide, make sure	functional	155
sostoit	consist of	Meronymy	260
sluzhit	serve to	functional	117
podraz-delâût	classify	classification	9
pozvol-âet	allow, enable	functional	115
Različaût	differen-tiate	classification	15
vklûčaet v sebâ	contain, com-prise	Meronymy	18
predstav-lâet soboj	is, cons-titutes	definition, explanation	56
ustanav-livaût	fix, mount	position indication, Meronymy	196
prednaz-načen	serve to, is meant to	functional	112
i drugie	and others	enumeration, classification	20

Table 1: Keywords of tentative knowledge patterns

Before the second extraction cycle, stop sentences were filtered out from the corpus, i. e.:

- incomplete sentences
- questions
- sentences beginning with stop words such as determiners and pronouns

These measures are essential for excluding sentences with anaphoric reference or single-case information which are responsible for a big share of noise in KRC extraction (cf. Meyer, 2001), but also for dealing with the particularities of internet text and unwanted pattern occurrences. By these measures, precision could be improved for some of our patterns. Sentences a) and b) are extracted example sentences:

a) Šassi avtomobilâ sostoit iz transmissii i hodovoj časti i mehanizmov upravleniâ.

The chassis of a car comprises the transmission, the frame and control equipment.

b) Sistema ohlaždeniâ služit dlâ otvoda izlišnego tepla ot detalej dvigatelâ, nagrevaûŝihsâ pri ego rabote.

The cooling system serves to remove excess heat from those parts of the motor which heat up during exploitation.

The results in the extraction experiments are still too variable to be considered final. Moreover, relevance decisions are not always straightforward. Questionable cases are erroneous sentences and associative contexts (cf. ISO, 2009). Another yet open problem is the extraction of lists containing classifications following introductory sentences on KPs such as *Različaût*, without which the KRC is worthless. In other cases, the extracted sentence is a KRC, but relates not to the target term, but to a closely related term. This is due to the absence of syntactic information, because of which KPs can be located at any position in the sentence, e. g. within dependency relations. In the experiment reported here these cases have, however, been evaluated as relevant KRCs. Table 2 presents Russian KPs that by now can be considered reliable:

KP	**Precision across experimental cycles**
sostoit	0,87-0,95
sluzhit	0,80-0,92
prednaz-načen	1,00
Različaût	1,00

Table 2: Reliable KPs

Other patterns such as *predstavlâet soboj* have stable results as well, but their occurrences in the studied corpus are too few to allow for final precision estimates.

4 Interpretation of results and future work

The outlined results shed light on two important shortcomings of pattern-based KRC extraction. The first one is data sparseness. Reliability estimations require large corpora that provide many pattern occurrences. This problem calls for a search strategy that uses the web as a corpus, otherwise dealing with very large local corpora and long lists of regular expressions will become intractable. There also is some hope that the problem of data quality mentioned in the previous section can be overcome by more data.

The second aspect is precision. It is clear that the KRCs described so far have an accidental element. The use of syntactic information in pattern creation may alleviate these shortcomings and provide a sound basis for the automated semantic analysis of extracted sentences by using semantic situation templates (Xu, 2007).

However, the advantage of the work described in this paper consists in its using light-weight methods. Acceptable results can be achieved for at least some of the tested patterns by means of a hand full of simple commands and tasks. In our view, this advantage of pattern-based approaches should not be given up easily. Our further work will therefore be directed at overcoming the difficulties mentioned. Moreover, bootstrapping methods will be tested by using reliable patterns established so far as seeds in order to identify more KPs. The developed method will be evaluated by means of an extraction task in a new domain and transferred to new languages such as German and Latvian.

Acknowledgments

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

References

Alain Auger, Caroline Barrière. 2008. Pattern-based approaches to semantic relation extraction. A state-of-the-art. *Terminology*, 14 (1): 1-19.

Alexander Maedche, Steffen Staab. 2000. Mining Ontologies from Text. *Lecture Notes in Computer Science*, 1937: 189-202.

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. *Lecture Notes in Computer Science*, 3060: 187-201.

Charles J. Fillmore. 1985. Frames and the Semantics of Understanding. Quaderni di Semantica VI (2): 222-254.

Christian Jacquemin, Didier Bourigault. 2003. Term Extraction and Automatic Indexing. Ruslan Mitkov (ed.). *The Oxford Handbook of Computational Linguistics*. Oxford University Press. Oxford: 599-615.

Eni Mustafaraj, Martin Hoof, Bernd Freisleben. 2006. Mining Diagnostic Text Reports by Learning to Annotate Knowledge Roles. Anne Kao, Steve Poteet (eds.). *Natural Language Processing and Text Mining*. Springer. London: 45-70.

Enrique Alfonseca, Maria Ruiz-Casado, Manabu Okumura, Pablo Castells. 2006. Towards Large-scale Non-taxonomic Relation Extraction: Estimating the Precision of Rote Extractors. 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, July 22, Sydney, Australia.

Eugene Agichtein, Luis Gravano. 2000. *Snowball*: Extracting Relations from Large Plain-Text Collections. 5th ACM International Conference on Digital Libraries 2000, June 2-7, San Antonio, USA.

Fei-Yu Xu. 2007. *Bootstrapping Relation Extraction from Semantic Seeds*. PhD Thesis. Saarland University Saarbrücken, Uszkoreit.

Gerardo Sierra, Rodrigo Alarcón, César Aguilar, Carme Bach. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology*, 14 (1): 74-98.

Ingrid Meyer. 2001. Extracting knowledge-rich contexts for terminography. Bourigault, Jacquemin, L'Homme (eds.): 279-302.

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). John Benjamins. Amsterdam, Philadelphia.

Khurshid Ahmad, Margaret Rogers. 2001. Corpus Linguistics and Terminology Extraction. Sue Ellen Wright, Gerhard Budin (eds.). *Handbook of Terminology Management.Vol 2: Application-Oriented Terminology Management*. John Benjamins. Amsterdam, Philadelphia: 725-760.

Marti A. Hearst. 1992. Automatic Acquisition of Hypernyms from Large Text Corpora. COLING 1992, August 23-28 1992, Nantes, France.

Nathalie Aussenac-Gilles, Brigitte Biébow, Sylvie Szulman. 2000. Revisiting Ontology Design: A Method Based on Corpus Analysis. *Lecture Notes in Computer Science*, 1937: 172-188.

Nathalie Aussenac-Gilles, Marie-Paule Jacques. 2006. Designing and Evaluating Patterns for Ontology Enrichment from Text. *Lecture Notes in Computer Science*, 4248: 158-165.

Paul Buitelaar, Daniel Olejnik, Michael Sintek. 2004. A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. 1st European Semantic Web Symposium May 10-12, Heraklion, Greece.

Robert Dubuc, Andy Lauriston. 1997. Terms and Contexts. Sue Ellen Wright, Gerhard Budin (eds.). *Handbook of Terminology Management. Vol. 1: Basic aspects of terminology management.* John Benjamins. Amsterdam, Philadelphia: 80-87.

Sergey Brin. 1998. Extracting Patterns and Relations from the World Wide Web. International Workshop on the Web and Databases, March 27-28 1998, Valencia, Spain.

Thomas Albert Sebeok. 1986. *Encyclopedic Dictionary of Semiotics. Vol. 1.* (Approaches to Semiotics 73). de Gruyter. Berlin, New York.

TTC project: Terminology Extraction, Translation Tools, Comparable Corpora. 2010. *Deliverable 2.1: First version of the crawler for comparable corpora*. The project is funded under the European Community's FP7/2007-2013, grant agreement n° 248005.

Véronique Malaisé, Pierre Zweigenbaum, Bruno Bachimont. 2005. Mining defining contexts to help structuring differential ontologies. *Terminology*, 11 (1): 21-53.