НА ПУТИ К АВТОМАТИЗИРОВАННОМУ ОБОГАЩЕНИЮ МНОГОЯЗЫЧНЫХ ТЕРМИНОЛОГИЧЕСКИХ БАЗ ДАННЫХ БОГАТЫМИ ЗНАНИЯМИ КОНТЕКСТАМИ

Шуман А. К. (anne.schumann@tilde.lv)

ООО «Тилде»/Венский университет, Рига/Вена, Латвия/Австрия

Ключевые слова: Корпусная лингвистика, богатые знаниями контексты, компьютерная терминография, русский язык, немецкий язык

TOWARDS THE AUTOMATED ENRICHMENT OF MULTILINGUAL TERMINOLOGY DATABASES WITH KNOWLEDGE-RICH CONTEXTS

Schumann A.-K. (anne.schumann@tilde.lv)

Tilde SIA/University of Vienna, Riga/Vienna, Latvia/Austria

This paper describes ongoing Phd thesis work dealing with the extraction of knowledge-rich contexts (KRCs) from specialized Russian and German text corpora for the semantic enrichment of terminological resources. In recent years, automatically extracted KRCs have been proposed as a means for deriving empirically grounded concept descriptions for terminography while maintaining the time and costs spent for the acquisition of such descriptions on a reasonable level. KRCs have been studied for a number of European languages ranging from English over French and Spanish to Catalan, however, not much effort has yet been put into widely spoken, but typologically different languages such as Russian or German. This paper, therefore, describes research efforts aiming at the extraction of KRCs in Russian and German for the purpose of termbase enrichment. Section 1 of this paper presents a brief introduction to KRC research and the motivation for this study. Section 2 gives an overview over related work. Section 3 describes the KnowPipe KRC extraction framework, whereas section 4 outlines ranking experiments with KnowPipe on Russian and German data. Section 5 summarizes the results and describes future work.

Key words: Corpus linguistics, knowledge-rich contexts, computer-aided terminography, Russian language, German language

1. Introduction

Definitions and explanations of concepts are an obligatory part of any termbase entry (ISO, 2009). However, there is no framework for the systematic enrichment of termbases with such content. In practice, semantic information is often added manually and unsystematically or omitted completely because of practical constraints. In this context, knowledge-rich context (KRC) extraction aims at identifying *semantic contexts* (as opposed to *linguistic contexts*) that provide *semantic information* about *concepts* (as opposed to *linguistic information* about *terms*) in text corpora and to feed the results of this process into a terminological resource. In other words, KRC extraction aims not only at providing examples of term occurrences, but sentences that provide additional knowledge about the concept to which a term refers. Research in this field, therefore, touches on aspects of terminology research that remain yet unresolved: Although the semantic types of contexts have been described in ISO 12620 (ISO, 2009), many terminological resources do not distinguish between different types of contexts and mainly restrict themselves to linguistic contexts and more or less informative usage examples. Often, however, contexts are completely omitted.

The aim of this study is to investigate, to which extent KRC extraction can contribute to the enrichment of a terminological resource with additional information by describing experiments conducted on two Russian corpora as well as on one German language corpus. In our experiments, the internet is used as a source of information since it is a primary means for finding information about terms and concepts for many professional translators and interpreters, and, in our view, a KRC extraction approach must therefore be able to deal with the quality of data found online in order to be applicable to real tasks.

2. Related Work

KRC extraction generally requires high precision, while specialized corpora from which KRCs can be extracted are typically small or must be crawled from online sources, a process that often outputs messy data. What is common to most studies in the field, therefore, is the fact that they employ a pattern-based method for KRC extraction. A systematic overview over pattern-based work is given by Auger and Barrière (2008). In most approaches, extraction patterns are acquired manually, but some groups (Condamines & Rebeyrolle, 2001; Halskov & Barrière, 2008) also devise a bootstrapping procedure for automated pattern acquisition similar to methods developed in information extraction (Xu, 2007). Seminal work for English was carried out by Pearson (1998) and Meyer (2001), and more recent work providing a contrastive linguistics perspective on English and French is Marshman (2007) and Marshman (2008). Recent studies for other languages are Feliu & Cabré (2002) for Catalan, Sierra et al. (2008) for Spanish, and Malaisé et al. (2005) for French. Sierra et al. (2008) employ 13 verbal patterns to extract instantiations of 3 previously defined types of definitions and implement a set of rules for filtering out irrelevant candidates. Malaisé et al. (2005) use a total of 42 verbal patterns for French as well as metalinguistic markers such as parentheses. For German, KRC extraction has not been studied, but Walter (2010) provides a detailed account on the related topic of extracting

definitions from court decisions. A recent study (Schumann, 2011) compares KRC extraction from German and Russian corpora.

As for the ranking of extraction output, Walter (2010) gives a detailed account of his experiments in the ranking of definition candidates using supervised machine learning techniques. The features used in his experiments can be divided into five groups:

- *Lexical*, such as boost words or stop words and features that are specific for legal language, such as subsumption signals
- *Referential*, such as anaphoric reference or definiteness of the definiendum
- *Structural*, such as the position of the definiendum relative to the definiens
- *Document-related*, such as the position of the definition candidate in the document and whether there are other candidates in its immediate context
- *Others*, such as sentence length or TF-IDF scores of terms in the sentence

Walter produces the best results using a linear regression algorithm. He also carries out experiments using the output of supervised classifiers such as Naïve Bayes or k-Nearest Neighbour as an additional feature in ranking.

3. Knowledge-Rich Context Extraction in Russian and German

3.1. Extraction Patterns

Previous studies of KRC extraction from Russian and German web corpora (Schumann, 2011) were based on a pattern-based extraction approach using mainly predicative Russian and German patterns. These patterns were combined either with target terms or morpho-syntactic term formation patterns to form regular expressions. Example 1 illustrates a lexical extraction trigger used in our Russian experiments and a valid KRC. The underlined term is the target term, the lexical extraction trigger is marked in bold. Example 2 illustrates a German KRC.

(1) ***Система охлаждения*** *служит для отвода излишнего тепла от деталей двигателя, нагревающихся при вго работе.*
[Translation: The ***cooling system*** *serves to remove excess heat from those parts of the engine that heap up during exploitation.]*

(2) *Das Verhältnis Energieertrag (Output) zu Input wird* ***Leistungszahl*** *genannt.*
[Translation: The relation between energy output and input is called ***coefficient of performance****.]*

Example 2 also illustrates that in KRC extraction from German text, it is necessary to deal with considerable syntactic complexity, since predicates used as lexical extraction triggers often consist of several surface words that are distributed across

the sentence forming long syntactic dependencies, into which the target term can be embedded — a problem that deserves further investigation.

Table 1 gives a simplified overview over the lexical patterns currently used by our extraction framework KnowPipe and assigns them to earlier defined semantic target relations (Schumann, 2011):

Table 1. Semantic relations and Russian extraction triggers

Relation	Explanation	Russian Patterns	German Patterns
Hyperonymy	Generic-Specific	Относить к, относиться к, включать в себя, классифицировать, различать, подразделять, разделять на, разделяться на, входить, составлять:, различать следующий, различать ... тип	Gehören zu, eine Form von ... sein, sein unter
Meronymy	Part-Whole	Состоять из, включать в себя, снабжать, снабдить, образовать, составлять:, оснащать, оснащаться, оснащенный	Bestehen aus, (sein|werden) (versehen | ausgestattet | ausgerüstet | bestückt), besitzen, verfügen über
Process	Temporal neighbourhood	Воздействовать, приводить, осуществлять	bewirken
Position	Spatial neighbourhood	Расположенный, располагать, устанавливать, устанавливаться	Sein (positioniert | angeordnet)
Causality	Cause-Effect	Обусловить, обуславливать, обеспечить, определяться	Ergeben sich aus, bewirken
Origin	Material or ideal origin	Состоять из	–

Relation	Explanation	Russian Patterns	German Patterns
Reference	General predication or definition	Представлять себя, называть, называться, определить, понимать, -это, так.называемый	(sein | werden) durch ... (charakterisiert | gekennzeichnet | beschrieben), werden genannt, (nennt | bezeichnet | unterscheidet) man, werden als ... bezeichnet, unter versteht man, werden (über | durch | mit) beschrieben, stellen dar, so.genannt
Function	Purpose or aim	Служить, позволять, предназначать, нужный для, применять, применяться	Stellen sich die Aufgabe, haben die (Aufgabe | Funktion), dienen (dazu|als), zu (nutzen | einsetzen | verwenden | benutzen)

The differences between the German and Russian pattern lists suggest that the latter are not yet definite, but need to be enlarged as soon as more information becomes available, e. g. about less-researched relations such as Origin.

3.2. KnowPipe

KnowPipe aims at providing a processing environment for multilingual KRC extraction using shallow as well as deep processing. In its current state of development, KnowPipe offers preprocessing tools for Russian and German text corpora, pattern-based KRC extraction as well as a ranking method based on shallow features for both languages.

Preprocessing consists of tokenization, removal of duplicate sentences, removal of stop sentences (e. g. incomplete sentences, questions) and lemmatization. The Perl Lingua::Sentence module[1] is used for tokenizing Russian and German corpora. TreeTagger (Schmid, 1994) is used for lemmatization. KRC candidates are extracted using the lexical patterns described in the previous section. They are then ranked directly according to the values outputted by a supervised machine learning algorithm, currently Naïve Bayes. The Naïve Bayes algorithm was chosen since it seemed to generalize best to different types of input data. The Perl Algorithm::NaiveBayes module[2] is used to carry out this process. It uses the following 13 features for ranking KRC candidates both in Russian and German:

---

[1] http://search.cpan.org/~achimru/Lingua-Sentence-1.00/lib/Lingua/Sentence.pm.

[2] http://search.cpan.org/~kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm.

Table 2. Shallow features used for ranking

Feature name	Explanation
Word tokens	The number of word tokens in the sentence.
Subscore	The normalized sum of the term relevance scores of terms constituting the subject.
Subpos	1 if the sentence starts with the subject, else 0.
Term score	The normalized sum of the term relevance scores scores of all other terms[3].
Nr. of terms	The number of terms in the sentence.
Position	1 if the subject is located before the extraction pattern, else 0.
Adjacent term	1 if there is a term directly adjacent to the extraction pattern, else 0.
Distance	An integer indicating the token distance between subject and pattern.
Negation	1 if the extraction pattern is preceded by a negation particle, else 0.
Boost words	1 if the pattern is preceded by a generalization signal, else 0.
Pattern score	A pattern reliability score based on baseline experiments (Schumann, 2011).
Stop words	Number of negative markers normalized by word tokens.
Definite Subject	1 if the subject is preceded by markers of definiteness or anaphora, else 0.

For Russian, we devised a heuristic that uses the rich annotation provided by the Russian TreeTagger tagset (Sharoff et al., 2008) and syntactic noun phrase formation patterns to identify noun phrases in nominative case. For German, we used the Parzu dependency parser (Sennrich et al., 2009) for analyzing sentence structure. Boost words are expressions such as *обычно, чаще всего* etc. that are indicative of a generalized statement for Russian and *üblicherweise, oft, häufig* etc. for German. Stop words, on the other hand, are expressions that are usually used for creating coherence, especially when the information provided in the sentence is single-case information and therefore not generalizable or makes reference to information that was given earlier in the text. Examples for stop words are: *иными словами, в итоге, поэтому, кроме того, например* etc for Russian and *dabei, aus diesem Grund, aber* etc. for German. The positional features are based on the hypothesis that even in a free word order language like Russian or German, KRCs favour canonical syntax over inverted

[3] An appropriate subsection of the Russian Internet Corpus (Sharoff, 2006) was used as a reference corpus in scoring. A search interface to this corpus is available here: http://corpus.leeds.ac.uk/ruscorpora.html. For German, we used a subsection of a 2011 newscrawl from the Leipzig corpora collection (Quasthoff et al., 2006), http://corpora.informatik.uni-leipzig.de/download.html.

structures. Moreover, a definite subject is treated as an indicator for single-case information which usually does not form a part of KRCs.

4. Experiments

4.1. Corpora and Experimental Setup

The performance of pattern-based KRC extraction from Russian and German text corpora has already been studied in Schumann (2011). For evaluating the performance of the ranking algorithm, we conducted experiments on the data outputted by this extraction step. For Russian, this data was extracted from two corpora, namely a small corpus dealing with the automotive domain and a larger corpus covering several engineering topics, as described in Schumann (2011). For German, KRC candidates extracted from a rather large, but noisy corpus covering the domains of electrical engineering and wind energy were used. For all corpora, a gold standard had been created earlier by means of manual annotation.

370 KRC candidates from the Russian automotive corpus and 709 KRC candidates from the larger Russian corpus were used for ranking. Since the ranking is carried out by means of a supervised learning algorithm, each data set had to be split into a training set in which each KRC candidate is marked as valid or invalid KRC and a test set on which the actual ranking is performed. On the Russian car corpus, 100 sentences were used for training and 270 for testing. On the Russian multidomain corpus, this relation was 300/409. The training sets were kept small to ensure the usability of the algorithm in a practical extraction task where typically not much data can be annotated manually. 322 frequently occurring terms were manually extracted from the gold standard for the car corpus. For the Russian multidomain corpus, the corresponding number was 372. These terms served as target terms — terms for which the extracted KRCs may supply additional information — in the feature annotation step. For German, the sample of KRC candidates extracted from the German multidomain corpus comprised 574 sentences. Out of these, 200 were manually annotated for training, the rest was used for testing. As for the Russian data sets, the earlier created gold standard was used as a reference. However, the number of manually extracted target terms was much higher for German than for Russian, namely 526.

4.2. Evaluation

For all three test samples Precision before and after ranking was calculated using the respective gold standards as reference. Table 3 gives an overview over the Precision values achieved on both Russian corpora as well as the German corpus before and after ranking for different Recall levels. Note that Recall was calculated with respect to the whole corpus, not just the test sample.

Table 3. Precision for different Recall levels before and after ranking

Recall levels	0,10	0,20	0,30	0,40
Russian car corpus				
Precision before ranking	0.49	0.45	0.47	0.51
Precision after ranking	0.97	0.85	0.73	0.49
Russian multidomain corpus				
Precision before ranking	0.38	0.40	0.35	–
Precision after ranking	0.89	0.66	0.38	–
German multidomain corpus				
Precision before ranking	0.47	0.4	0.33	–
Precision after ranking	0.8	0.69	0.45	–

5. Discussion and Future Work

The results outlined in the previous section are encouraging in the sense that the ranking algorithm seems to support the selection of valid KRCs: Valid KRC candidates are moved to the top of the sample, whereas invalid candidates are moved to the bottom, which produces relatively high Precision in the top-n segment and lower Precision for the bottom of the test sample, allowing for an easier selection of valid KRCs at the end of the process. Since no language-specific features are used, the ranking algorithm also generalizes to German. However, the results obtained on the German data still need to be considered preliminary as no experiments with a second corpus have been conducted yet. In terms of overall performance — especially with respect to a real-world database enrichment task — several improvements still need to be undertaken, including the improvement of overall Precision which is likely to result in a better performance also during ranking. In the future, we plan to conduct experiments on using deeper linguistic knowledge, e. g. syntactic information, for creating more powerful — and thus more precise — extraction patterns. This seems especially relevant to the case of German, however, we believe that the use of more linguistic information will improve the performance of both the extraction step and the ranking algorithm also on the Russian data. Recall still calls for further improvement and experiments need to be conducted on larger datasets. For termbase enrichment, means need to be devised for defining the target term of each KRC candidate, e. g. for establishing an explicit relation between an extraction pattern and a term that is being explained by the phrases introduced by the pattern. There is reason to believe that richer linguistic information will also be helpful in this task.

Acknowledgement

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

References

1. *Auger, A., Barrière, C.* (2008), Pattern-based approaches to semantic relation extraction. Terminology, Vol. 14 (1), pp. 1–19.
2. *Condamines, A., Rebeyrolle, J.* (2001), Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB). In D. Bourigault, C. Jacquemin, M.-C. L'Homme (Eds.), Recent Advances in Computational Terminology, Amsterdam/Philadelphia, John Benjamins, pp. 127–148.
3. *Feliu, J., Cabré, M.* Conceptual relations in specialized texts: new typology and an extraction system proposal. Proceedings of TKE, Nancy, 2002, pp. 45–49.
4. *Halskov, J., Barrière, C.* (2008), Web-based extraction of semantic relation instances for terminology work. Terminology, Vol. 14 (1), pp. 20–44.
5. *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. Geneva, ISO.
6. *Malaisé, V., Zweigenbaum, P., Bachimont, B.* (2005), Mining defining contexts to help structuring differential ontologies. Terminology, Vol. 11 (1), pp. 21–53.
7. *Marshman, E.* (2007), Towards strategies for processing relationsips between multiple relation participants in knowledge patterns. An analysis in English and French. Terminology, Vol. 13 (1), pp. 1–34.
8. *Marshman, E.* (2008), Expressions of uncertainty in candidate knowledge-rich contexts. A comparison in English and French specialized texts. Terminology, Vol. 14 (1), pp. 124–151.
9. *Meyer, I.* (2001), Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework. In Recent Advances in Computational Terminology, pp. 279–302.
10. *Pearson, J.* (1998), Terms in Context. (Studies in Corpus Linguistics 1). Amsterdam/Philadelphia, John Benjamins.
11. *Quasthoff, U., Richter, M., Biemann, C.,* Corpus Portal for Search in Monolingual Corpora, Proceedings of LREC, Genoa, 2006, pp. 1799–1802.
12. *Schmid, H.,* Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing, Manchester, 1994, pp. 44–49.
13. *Schumann, A.-K.,* A Bilingual Study of Knowledge-Rich Context Extraction in Russian and German. Proceedings of the Fifth Language & Technology Conference, Poznan, 2011, pp. 516–520.
14. *Sennrich, R., Schneider, G., Volk, M., Warin, M.,* A New Hybrid Dependency Parser for German. Proceedings of GSCL Conference, Potsdam, 2008.
15. *Sharoff, S.* (2006), Creating general-purpose corpora using automated search engine queries. In Baroni, M., Bernardini, S. (Eds.), WaCky! Working papers on the Web as Corpus. Bologna, Gedit.
16. *Sharoff, S., Kopotev, M., Erjavec, T., Feldmann, A., Divjak, S.,* Designing and evaluating Russian tagsets. Proceedings of LREC, Marrakech, 2008.

17. *Sierra, G., Alarcón, R., Aguilar, C., Bach, C.* (2008), Definitional verbal patterns for semantic relation extraction. Terminology, Vol. 14 (1), pp. 74–98.
18. *Walter, S.* (2010), Definitionsextraktion aus Urteilstexten. PhD thesis in Computational Linguistics. Saarland University Saarbrücken.
19. *Xu, F.-Y.* (2007), Bootstrapping Relation Extraction from Semantic Seeds. PhD thesis in Computational Linguistics. Saarland University Saarbrücken.