Towards the Automated Enrichment of Multilingual Terminology Databases with Knowledge-Rich Contexts – Experiments with Russian EuroTermBank Data

Anne-Kathrin Schumann

University of Vienna / Tilde
anne.schumann@tilde.lv

Abstract. Although knowledge-rich context (KRC) extraction has received a lot of attention, to our knowledge few attempts at directly feeding KRCs into a terminological resource have been undertaken. The aim of this study, therefore, is to investigate to which extent pattern-based KRC extraction can be useful for the enrichment of terminological resources. The paper describes experiments aiming at the enrichment of a multilingual term bank, namely EuroTermBank, with KRCs extracted from Russian language web corpora. The contexts are extracted using a simple pattern-based method and then ranked by means of a supervised machine learning algorithm. The internet is used as a source of information since it is a primary means for finding information about terms and concepts for many language professionals, and 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.

Keywords: computer-aided terminography, knowledge-rich contexts, web as corpus, Russian language, multilingual terminology databases

1 Introduction and Related Work

In recent years, knowledge-rich context (KRC) extraction has been put forward as a means for enriching existing multilingual terminology resources with concept definitions and explanations while keeping the acquisition effort on a justifiable level. KRCs can be defined as follows (see [10], [13]):

Definition 1. Knowledge-rich contexts are 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 understand the concept in question.

KRC extraction aims at identifying 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. It therefore touches upon

aspects of terminology research that remain yet unresolved: although the different types of contexts have been described in ISO 12620 ([6]), many terminological resources do not distinguish between various context types and often restrict themselves to linguistic contexts and more or less informative usage examples. In other cases, contexts are completely omitted.

The extraction of KRCs has been actively researched for several languages. Seminal work for English was carried out by [11] and [10], and recent studies providing a contrastive perspective on English and French are [8] and [9]. Recent work on other languages are [4] for Catalan, [16] for Spanish, and [7] for French. [17] studies the topic of definition extraction from German court decisions, whereas [13] gives a first evaluation of KRC extraction patterns for Russian and German.

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 many studies in the field, therefore, is the fact that they employ a pattern-based method. A systematic overview over pattern-based work is given by [1]. Often, extraction patterns are acquired manually, but some groups ([2], see [5]) also devise a bootstrapping procedure for automated pattern acquisition similar to methods developed in information extraction ([18]).

As for the ranking of extraction output, [17] 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

2 Towards the Enrichment of EuroTermBank

2.1 EuroTermBank

EuroTermBank[1] ([12]) is a multilingual term bank that was released in 2007. More specifically, it is a terminology repository binding together specialized terminology collections in 27 European languages. The terminology collections represented in EuroTermBank (ETB) consist of electronic collections contributed from various partners as well as digitalized versions of print dictionaries. Special attention was paid to providing resources for small and under-resourced languages especially from the new EU member-states, such as the Baltic languages. In terms of entries, the 5 best-resourced languages in EuroTermbank are English, Russian, German, Latvian, and Polish (in this order).

[1] http://eurotermbank.com/.

2.2 Knowledge-Rich Context Extraction in Russian

Previous studies of KRC extraction from Russian web corpora ([13]) were based on a pattern-based extraction approach using 47 mainly predicative Russian patterns. These patterns had been combined either with target terms or morpho-syntactic term formation patterns to form regular expressions. In our present experiments, we used a similar approach, but extraction was applied to lemmatized text in order to facilitate the process and extraction patterns were used without any kind of term representation. Example 1 illustrates a lexical extraction trigger and a valid KRC extracted in the course of our experiments. The underlined term is an ETB target term, whereas the lexical extraction trigger is marked in bold.

Example 1. Эстафетная палочка **представляет собой** цельную, гладкую, полую трубку, круглую в сечении, сделанную из дерева, металла или другого твердого материала.
*(The relay baton **is** a one-piece, smooth, hollow, and round tube made from wood, metal or another hard material.)*

Semantic relations are elementary building blocks of KRCs. We therefore devised a typology of semantic target relations that make up a valid KRC. Table 1 gives an overview over these relations along with examples of lexical extraction triggers:

Table 1. Semantic relations and Russian extraction triggers

Relation	Explanation	Patterns	Translation
Hyperonymy	Generic-Specific	Относить к, включать в себя	Belong to, include
Meronymy	Part-Whole	Состоять из	Consist of
Process	Temporal neighbourhood	Воздействовать	Act upon
Position	Spatial neighbourhood	Располагать	Locate
Causality	Cause-Effect	Обусловить	Determine
Origin	Material or ideal origin	Состоять из	Is made of
Reference	General predication or definition	Представлять себя, называть	Is, call
Function	Purpose or aim	Служить, позволять	Serve, allow

2.3 Ranking

KRC candidates are extracted using the patterns described in the previous section. They are then ranked directly according to the values outputted by a Naïve Bayes classification algorithm. The Perl Algorithm::NaiveBayes module[2] is used to carry out this procedure based on the following 13 features:

[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 of all other terms.
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	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 estimate.
Stop words	Number of negative markers normalized by word tokens.
Definite Subject	1 if the subject is preceded by markers of definiteness or anaphora.

In order to identify the subject of a sentence, a heuristic using the rich annotation provided by the Russian TreeTagger tagset ([15]) and syntactic noun phrase formation patterns as observed in our corpus was devised. As for the term scoring method, we achieved the best results not by using a classical TF-IDF score, but a slightly modified score that takes into account relative term frequency as well as the occurrence of the target term in the extraction corpus and a reference corpus[3]. This score outputs values higher than zero for all terms that occur in at least one of the corpora and always ranks frequent terms higher than less frequent terms, which corresponds to the hypothesis that the existence of a valid KRC is more likely, if the target term is highly frequent. The development and adaptation of the best term scoring method will be further studied in future experiments. The positional features in our ranking scheme are based on the hypothesis that even in a language with relatively free word order such as Russian sentences that contain definitional information favour a regular word order. Boost words are generalization signals such as *часто* (often) or *обычно* (usually), whereas stop words include outdated language such as *СССР* (USSR) and *советский* (soviet).

[3] The Russian Internet Corpus ([14]) was used as a reference corpus. A search interface to this corpus is available here: http://corpus.leeds.ac.uk/ruscorpora.html.

3 Experiments on Enriching EuroTermBank with Knowledge-Rich Contexts

3.1 Resource Selection and Corpus creation

We selected a rather small ETB resource, namely the athletics domain. For Russian, this domain comprises 665 entries from which the target terms were harvested. The final term list has 667 target terms. Some of these terms are verb phrases, others are rather generic terms such as *скорая помощь* (first aid) and „*ветер*“ (wind), or polysemic such as *построение* (which often means “formation” or “construction”, but in ETB’s athletics domain is translated to English as “line-up”) and “*Нет!*” (No!), which is given as a synonym for *прыжок не засчитан* (the jump was not counted).
We used some of the target terms harvested from ETB as seeds in a corpus crawling process. The corpus crawler was Babouk ([3]). However, the term list obtained from ETB had to be cleaned in order to remove the following shortcomings:

- Some entries contain synomyms or near synonyms separated by commas. In such cases, the synonyms were treated as two separate target terms.
- If very general terms are fed into Babouk, the obtained corpus is likely to contain a high percentage of out-of-domain texts, since the seed terms are polysemic. Therefore, most unigrams were removed from the seed list.

Moreover, for each seed term, more than one word form was supplied in order to improve the performance of Babouk. The crawling process had to be repeated several times. The resulting corpus has 517.266 running words and 28.448 sentences after cleaning. Table 3 gives an overview over the 10 most frequent ETB term concordances in the corpus.

Table 3. Overview over 10 most frequent ETB terms in corpus

Term	Translation	Count	Term	Translation	Count
бег	running	4254	подготовка	preparation	1353
техника	technique	1648	дистанция	distance	1336
соревоноание	competition	1467	прыжок	jump	1106
скорость	speed	1396	шаг	step	998
спортсмен	athlete	1229	выносливость	endurance	843

Out of our initial 667 terms, 420 were found in the corpus, and out of those, 209 had at least 10 and 102 at least 50 concordances.

3.2 Experimental Setup and Results

KRC extraction for term bank enrichment besides filtering KRC candidates from unseen data includes two more tasks, namely the attribution of the explanation provided by the KRC candidate to a specific target term and the filtering of KRCs that are not related to any of the relevant target terms.

To test the performance of our current method on these tasks, we extracted KRC candidates from the sports corpus. This process outputted 3068 KRC candidates. Unlike the experiments described in [13] no morpho-syntactic target term representation was used in this step, resulting in a very simple extraction method and a large amount of data. On this data, we conducted two experiments. In the first setting, ranking was performed only on those KRC candidates, for which the feature extraction step revealed a target term in subject position. In a second experiment, we applied the ranking algorithm to all KRC candidates that matched at least one term. Since the ranking algorithm is based on supervised learning, each data set had to be split into a training and a test set. Table 4 gives an overview over the data sets.

Table 4. Datasets used in experiments

Setting	Overall size of data set	Size of training set	Size of test set
Subject setting	521 KRC candidates	100	421
Term setting	1813 KRC candidates	300	1513

The ranking algorithm was applied to select valid KRCs from the datasets and simple heuristics were devised in order to find the target term of each KRC candidate: in the subject setting, the subject of each sentence was set to be the target term, whereas in the term setting a cascaded procedure for target term selection was applied:

- If there was a term in subject position, this term was set to be the target term.
- Otherwise, a term directly adjacent to the extraction pattern – if applicable – was set to be the target term.
- If none of these conditions was met, the first matching term in the sentence was set to be the target term.

Results were manually evaluated by picking and evaluating the highest ranked sentence for each term. Sentences with very low ranks were not evaluated. For target terms that are verb phrases, a relaxed setting was applied by accepting sentences that contain valid collocations, since it is yet unclear how the concept of KRCs can be applied to verbs. Table 5 presents the results.

Table 5. Results obtained in two experimental settings

Setting	Number of evaluated sentences	Unique KRC candidates for ETB target terms	Correct unique KRCs	Precision of attribution of KRC candidate to target term
Subject setting	407	82	57	0.96
Term setting	1504	197	112	0.91

4 Discussion and Future Work

The results of our experiments suggest that even in a very relaxed extraction setting, the current KRC extraction method achieves only limited coverage. More specifically, only for roughly 18% of our initial 667 target terms and 28% of all ETB terms in the corpus unique valid KRCs could be found including KRCs found during the manual annotation of the training sets. For higher recall, the pattern-based method may need to be supplemented by other methods that might be applied to the data in an iterative fashion. The systematic use of term variants may also help to retrieve more relevant contexts from the corpus.

The fact that the more relaxed term setting outperforms the subject setting in terms of coverage suggests that future research efforts should concentrate on the use of more linguistic information for higher precision and better ranking results to support the selection of valid candidates: In our view, the improvement of the current method by applying deeper linguistic knowledge such as syntactic information and making wider use of morphology will help establish a link between an ETB term and a lexical extraction trigger, thus eliminating noise and resulting in better ranking and target term selection. Other aspects that deserve to be mentioned are term-inherent polysemy affecting the process starting already upon corpus crawling. Moreover, more sophisticated processing such as the filtering of proper names and ambiguity resolution for polysemic terms may improve results.

Last but not least, the results outlined in this paper show that KRC extraction can be just one means of term bank enrichment: the current method deals but weakly with terms that are verbs and verb phrases and other kinds of information, e.g. collocations, may indeed be the better choice for this particular kind of terms.

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

References

[1] Auger, A., Barrière, C.: Pattern-based approaches to semantic relation extraction. Terminology. 14 (1), 1-19 (2008)

[2] Condamines, A., Rebeyrolle, J.: Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB). In: Bourigault, D., Jacquemin, C., L'Homme, M.-C. (eds.) Recent Advances in Computational Terminology, pp. 127-148. John Benjamins, Amsterdam/Philadelphia (2001)

[3] De Groc, C.: Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In: IEEE/WIC/ACM International Conference on Web Intelligence (2011)

[4] Feliu, J., Cabré, M.: Conceptual relations in specialized texts: new typology and an extraction system proposal. In: Proceedings of TKE 2002, pp. 45-49. INRIA, Nancy (2002)
[5] Halskov, J., Barrière, C.: Web-based extraction of semantic relation instances for terminology work. Terminology. 14 (1), 20-44 (2008)
[6] International Organization for Standardization. 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 (2009)
[7] Malaisé, V., Zweigenbaum, P., Bachimont, B.: Mining defining contexts to help structuring differential ontologies. Terminology. 11 (1), 21-53 (2005)
[8] Marshman, E.: Towards strategies for processing relationsips between multiple relation participants in knowledge patterns. An analysis in English and French. Terminology. 13 (1), 1-34 (2007)
[9] Marshman, E.: Expressions of uncertainty in candidate knowledge-rich contexts. A comparison in English and French specialized texts. Terminology. 14 (1), 124-151 (2008)
[10] Meyer, I.: Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework. In: Bourigault, Jacquemin, L'Homme (eds.), pp. 279-302 (2001)
[11] Pearson, J.: Terms in Context. (Studies in Corpus Linguistics 1). John Benjamins, Amsterdam/Philadelphia (1998)
[12] Rirdance, S., Vasiļjevs, A. (eds.): Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project. Tilde, Riga (2006)
[13] Schumann, A.-K.: A Bilingual Study of Knowledge-Rich Context Extraction in Russian and German. In: Proceedings of the Fifth Language & Technology Conference, pp. 516-520. Fundacja Uniwersytetu im. A. Mickiewicza, Poznan (2011)
[14] Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In: Baroni, M., Bernardini, S. (eds.), *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna (2006)
[15] Sharoff, S., Kopotev, M., Erjavec, T., Feldmann, A., Divjak, S.: Designing and evaluating Russian tagsets. In: *Proceedings of LREC* (2008)
[16] Sierra, G., Alarcón, R., Aguilar, C., Bach, C.: Definitional verbal patterns for semantic relation extraction. Terminology. 14 (1), 74-98 (2008)
[17] Walter, S.: Definitionsextraktion aus Urteilstexten. PhD thesis in Computational Linguistics. Saarland University Saarbrücken (2010)
[18] Xu, F.-Y.: Bootstrapping Relation Extraction from Semantic Seeds. PhD thesis in Computational Linguistics. Saarland University Saarbrücken (2007)