Knowledge-Rich Context Candidate Extraction and Ranking with KnowPipe

Anne-Kathrin Schumann
University of Vienna
Vienna, Austria
Tilde SIA
Rīga, Latvia
E-mail: anne.schumann@tilde.lv

Abstract

This paper presents ongoing Phd thesis work dealing with the extraction of knowledge-rich contexts from text corpora for terminographic purposes. Although notable progress in the field has been made over recent years, there is yet no methodology or integrated workflow that is able to deal with multiple, typologically different languages and different domains, and that can be handled by non-expert users. Moreover, while a lot of work has been carried out to research the KRC extraction step, the selection and further analysis of results still involves considerable manual work. In this view, the aim of this paper is two-fold. Firstly, the paper presents a ranking algorithm geared at supporting the selection of high-quality contexts once the extraction has been finished and describes ranking experiments with Russian context candidates. Secondly, it presents the KnowPipe framework for context extraction: KnowPipe aims at providing a processing environment that allows users to extract knowledge-rich contexts from text corpora in different languages using shallow and deep processing techniques. In its current state of development, KnowPipe provides facilities for preprocessing Russian and German text corpora, for pattern-based knowledge-rich context extraction from these corpora using shallow analysis as well as tools for ranking Russian context candidates.

Keywords: computer-aided terminography, knowledge-rich contexts, corpus-based terminology

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 in the form of definitions or explanations is often added manually and unsystematically or omitted completely because of practical constraints. In this context, the large-scale extraction of knowledge-rich contexts (KRCs) from corpora has been proposed as a means for enriching terminological resources with definitions and explanations, while keeping the effort on a justifiable level. In our work, the following definition of KRCs (Schumann, 2011; Meyer, 2001) is used:

(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.

This definition excludes dictionary definitions of concepts and puts special focus on the usefulness of the information to the end user. As can be seen from the above, semantic relations play a crucial role for the description of KRCs since they describe the content elements that are relevant to the description of a concept. It is therefore necessary to arrive at a workable definition of those semantic relations that are relevant to KRCs. For doing so, we carried out a comparison of several typologies of semantic relations (Schumann, 2011) and defined a set of semantic target relations that make up a valid KRC:

- Hyperonymy
- Meronymy
- Process
- Position
- Causality
- Origin
- Reference
- Function

The first two of these relations are well-known and correspond to the generic and partitive relations in ISO 12620: 2009 (ISO, 2009)[1], whereas *Process*, *Position*, and *Causality* correspond to the temporal, sequential, and causal relations in the same norm. The relation *Origin* is supposed to describe the material or ideal origin of the object to which a concept refers. *Reference* relates to simple predications that cannot be grouped under Hyperonymy. *Function* is an important semantic relation (see Murphy, 2003) and therefore also added to the list of target relations. We believe that by applying the above definition and inventory of target relations, we can define validity criteria for distinguishing valid KRCs as those given in (2) and (3) from invalid material:

(2)

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

[1] Cited from ISOcat: http://www.isocat.org/interface/index.html, accessed March 8, 2012.

(3)

Das Blattwinkelverstellsystem hat die Aufgabe, die Blätter in der richtigen Position genau einzustellen, aber auch die Blätter im Notfall in eine sichere Position zu bringen.
[Translation: The rotator control system's task is to accurately fix the rotors in the correct position and to move them to safety position in case of emergency.]

Russian, the language we mainly deal with in this paper, belongs to the Slavic language family and is characterized by free word order and rich morphology. Russian has six grammatical cases in singular and plural as well as some other interesting linguistic features like verbal aspect and *Aktionsarten*. Many of these features distinguish Russian from the languages that up to now have been in the focus of KRC extraction research. It seems therefore reasonable to study how well existing approaches generalize to a typologically different language like Russian and which methods need to be applied in order to grasp the particularities of this language and improve extraction results. Another open issue is the question how valid KRCs can be selected from overall extraction results. In this paper, we propose a ranking approach for this task that makes use of shallow features and describe experiments with Russian data.

2. Related Work

The extraction of KRCs has been actively researched in recent years. 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. 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. 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 all of the above-mentioned studies, 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 & Barrière (2008). In the cited 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).
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 four 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

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. The KnowPipe Framework

A schematic overview over KnowPipe is given in Figure 1. The framework can be roughly divided into four parts, namely preprocessing, pattern matching, ranking, and retrieval of original sentences. The framework combines language-independent and language-specific tools that were implemented in Perl. Preprocessing consists of sentence splitting, removal of duplicate and stop sentences, and lemmatization. The Perl Lingua::Sentence module[2] is used for splitting German corpora and rules were added for dealing with Russian. Stop sentences are sentences such as questions or incomplete sentences that are unlikely to be valid KRCs. TreeTagger (Schmid, 1994) is used for lemmatizing Russian and German text. The result is a data file that juxtaposes the sentences from the corpus with their lemmatized counterparts together with a common index.
For finding KRC candidates, a simple pattern matching approach is used for both Russian and German (Schumann, 2011). Patterns used in this step typically consist of a morpho-syntactic term formation pattern and a lexical extraction trigger, often a predicate. Previous results suggest that simple pattern-based methods give encouraging results on Russian and seem to be flexible enough to capture free word order, but on German, a language with long syntactic dependencies, syntactically informed methods are likely to provide more satisfactory output.
In the current state of development, a ranking step has been included for sorting the KRC candidates extracted from Russian corpora according to their quality. This seems reasonable, as the results of the extraction step are not yet satisfactory in terms of precision. A feature annotator was built for annotating linguistic features in each KRC candidate. The Perl Algorithm::NaiveBayes module[3] is used to carry out the ranking.
In the last step, the original counterparts of the extracted sentences are retrieved from the data file by means of their index and ordered according to their ranking value.

[2] http://search.cpan.org/~achimru/Lingua-Sentence-1.00/lib/Lingua/Sentence.pm, accessed March 9, 2012.
[3] http://search.cpan.org/~kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm, accessed March 9, 2012.

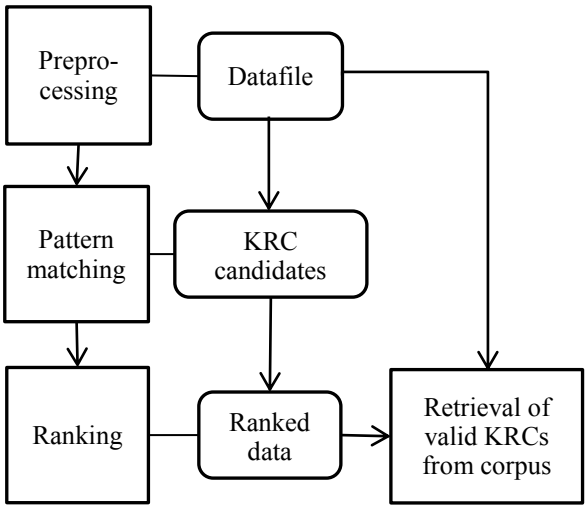

Figure 1: The KnowPipe Framework

4. Ranking experiments

4.1 Preparatory work

We conducted experiments on ranking KRC candidates using the output of the extraction experiments described in Schumann (2011) for Russian. The KRC candidates were extracted from two earlier collected Russian web corpora, namely a small corpus dealing with automotive texts and a larger, but lower quality corpus covering several topics ranging from electrical engineering to energy supply. For both corpora, a gold standard had been created by manual annotation of target KRCs.

In our own ranking experiments, we used similar features to those introduced by Walter (2010). However, the use of document-related features is not straightforward on web corpora, so we did not make use of this feature type. In the current implementation of KnowPipe, 13 linguistic features are used for ranking:

- *Word tokens* is the number of word tokens for each sentence.
- *Subscore* is a numeric value calculated as the sum of the TF-IDF score[4] of all terms that are part of the subject and normalised by the number of terms that participate in the subject.
- *Subpos* is a flag that indicates whether the sentence starts directly with the subject or not.
- The *Term score* is the normalized sum of the TF-IDF scores of all terms besides the subject.
- *Position* is a flag that indicates whether the subject of the sentence is located before the pattern that triggered extraction.
- *Adjacent term* is a flag that indicates whether there is a term directly adjacent to the knowledge pattern.
- *Distance* measures the distance between subject and knowledge pattern.
- *Boost words* flags whether the pattern is preceded by a lexical generalization signal such as *иными словами* (in other words), *в итоге* (consequently), *поэтому* (because of that), *например* (for example) etc.
- The *Pattern score* is a reliability estimation for each extraction pattern based on the experiments described in Schumann (2011).
- *Stop words* is the number of negative lexical markers (such as anaphora) divided by the number of word tokens.
- *Definite Subject* is a flag that indicates whether the subject is preceded by a marker of definiteness, based on the hypothesis that definiteness relates to single case information.

The positional features in our ranking scheme are based on the hypothesis that even in free word order languages KRCs favour a canonical word over an inverted order.

In the absence of a freely available syntactic parser for Russian, KnowPipe uses the rich annotation provided by the Russian TreeTagger tagset[5] in combination with noun phrase formation patterns for finding noun phrases in nominative case. This heuristic serves to identify the subject of each KRC candidate.

370 KRC candidates from the automotive corpus and 709 KRC candidates from the larger corpus were used for ranking experiments. On the car corpus, 100 sentences were used for training and 270 for testing. On the 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 terms were manually extracted from the gold standard for the car corpus. For the multidomain corpus, the corresponding number was 372. These terms served as target terms in the feature annotation step.

The earlier created gold standard was used as a reference for determining whether a candidate is valid or not and this information was used as the dependent variable in training and testing the Naïve Bayes algorithm. Overall 335 target KRCs were annotated in the car corpus and 422 target KRCs in the multidomain corpus. The Naïve Bayes classifier was chosen since it seemed to generalize best to our type of data. Logistic regression gave encouraging results on the car corpus, but failed on the multidomain corpus. The value outputted by the algorithm is not used for classifying the KRC candidates since it seems more reasonable to keep all candidates and rank them for manual inspection or further processing. Therefore, we decided to rank the candidates according to the value outputted by the classifier.

4.2 Results and Evaluation

For evaluation, we calculated Precision at different Recall levels on both samples before and after ranking. Table 1 gives an overview over the values achieved in this experiment. Please note that Recall was calculated in relation to the whole corpus, not just the test sample.

[4] We used appropriate subsets of the Russian internet corpus (Sharoff, 2006) as a reference in scoring.

[5] http://corpus.leeds.ac.uk/mocky/, accessed March 13, 2012.

Recall levels	0,10	0,20	0,30	0,40
Car corpus				
Precision before ranking	0.49	0.45	0.47	0.51
Precision after ranking	0.97	0.85	0.73	0.49
Multidomain corpus				
Precision before ranking	0.38	0.40	0.35	-
Precision after ranking	0.89	0.66	0.38	-

Table 1: Precision for different Recall levels before and after ranking

The scores achieved in this experiment suggest that the algorithm is successful in determining which KRC candidates should be ranked higher than others. The Precision values indicate that after ranking in both output samples the top n candidates are valid KRCs, whereas the invalid ones are moved to the bottom section of the sample. Since the algorithm does not make use of domain-specific features, it generalizes to the multi-domain corpus. Figure 2 gives a precision graph comparison for both the car and the multidomain corpus before and after ranking with respect to the number of retrieved valid KRCs. The upper graphs indicate not interpolated precision values after ranking and the graphs beneath visualize the not interpolated precision values before ranking.

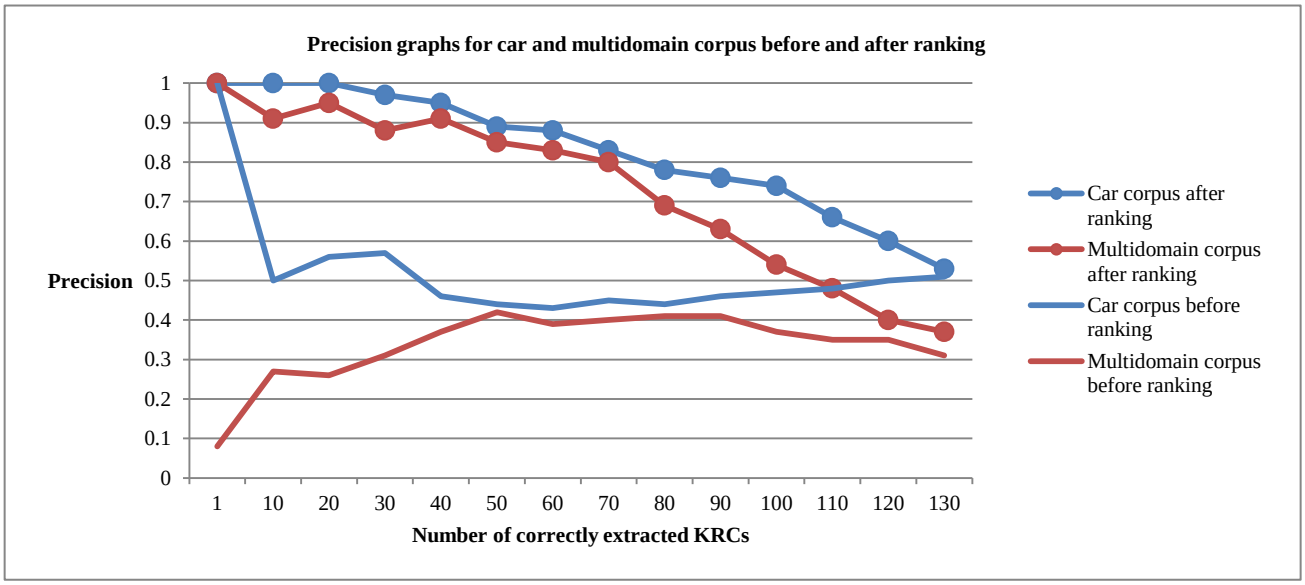

Figure 2: Precision graphs for car and multidomain corpora before and after ranking

5. Discussion and Future Work

The results presented in the previous section can be considered encouraging, however, it still needs to be shown how the ranking approach generalizes to new data, e. g. data from new languages (such as German) or domains. Moreover, a real world use case, e. g. a termbase enrichment scenario, still needs to be tested. A major concern of future work will be with the integration of deep processing techniques into the framework since we believe that refined patterns containing more linguistic information will provide better results and help improve the ranking algorithm especially in the case of German. Another focus will be with experiments on the automated acquisition of new extraction patterns for improved recall.

Acknowledgements

The research described in this paper was funded under the CLARA project (FP7/2007-2013), grant agreement n° 238405. Cordial thanks also go to my colleague Roberts Rozis for his many ideas and suggestions.

References

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

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.

Feliu, J., Cabré, M. (2002). Conceptual relations in specialized texts: new typology and an extraction system proposal. *Proceedings of TKE 2002*, Nancy, France: INRIA, pp. 45-49.

Halskov, J., Barrière, C. (2008). Web-based extraction of semantic relation instances for terminology work. *Terminology*, 14 (1), pp. 20-44.

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.
Malaisé, V., Zweigenbaum, P., Bachimont, B. (2005). Mining defining contexts to help structuring differential ontologies. *Terminology*, 11 (1), pp. 21-53.
Marshman, E. (2007). Towards strategies for processing relationsips between multiple relation participants in knowledge patterns. An analysis in English and French. *Terminology*, 13 (1), pp. 1-34.
Marshman, E. (2008). Expressions of uncertainty in candidate knowledge-rich contexts. A comparison in English and French specialized texts. *Terminology*, 14 (1), pp. 124-151.
Meyer, I. (2001). Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework. In Bourigault, Jacquemin, L'Homme (Eds.), pp. 279-302.
Murphy, M. (2003). Semantic Relations and the Lexicon. Antonymy, Synonymy, and Other Paradigms. Cambridge: Cambridge University Press.
Pearson, J. (1998). Terms in Context. (Studies in Corpus Linguistics 1). Amsterdam/Philadelphia: John Benjamins.
Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, England, pp. 44-49.
Schumann, A.-K. (2011). A Bilingual Study of Knowledge-Rich Context Extraction in Russian and German. *Proceedings of the Fifth Language & Technology Conference*. Poznan, Poland: Fundacja Uniwersytetu im. A. Mickiewicza, pp. 516-520.
Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni, S. Bernardini (Eds.): *WaCky! Working Papers on the Web as Corpus*. Bologna: Gedit, pp. 63-98.
Sierra, G., Alarcón, R., Aguilar, C., Bach, C. (2008). Definitional verbal patterns for semantic relation extraction. *Terminology,* 14 (1), pp. 74-98.
Walter, S. (2010). Definitionsextraktion aus Urteilstexten. PhD thesis in Computational Linguistics. Saarland University Saarbrücken.
Xu, F.-Y. (2007). Bootstrapping Relation Extraction from Semantic Seeds. PhD thesis in Computational Linguistics. Saarland University Saarbrücken.