Topic Extraction from Scientific Literature for Competency Management

Paul Buitelaar, Thomas Eigner

DFKI GmbH
Language Technology Lab & Competence Center Semantic Web
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
paulb@dfki.de

Abstract We describe an approach towards automatic, dynamic and time-critical support for competency management and expertise search through topic extraction from scientific publications. In the use case we present, we focus on the automatic extraction of scientific topics and technologies from publicly available publications using web sites like Google Scholar. We discuss an experiment for our own organization, DFKI, as example of a knowledge organization. The paper presents evaluation results over a sample of 48 DFKI researchers that responded to our request for a-posteriori evaluation of automatically extracted topics. The results of this evaluation are encouraging and provided us with useful feedback for further improving our methods. The extracted topics can be organized in an association network that can be used further to analyze how competencies are interconnected, thereby enabling also a better exchange of expertise and competence between researchers.

1 Introduction

Competency management, the identification and management of experts on and their knowledge in certain competency areas, is a growing area of research as knowledge has become a central factor in achieving commercial success. It is of fundamental importance for any organization to keep up-to-date with the competencies it covers, in the form of experts among its work force. Identification of experts will be based mostly on recruitment information, but this is not sufficient as competency coverage (competencies of interest to the organization) and structure (interconnections between competencies) change rapidly over time. The automatic identification of competency coverage and structure, e.g. from publications, is therefore of increasing importance, as this allows for a sustainable, dynamic and time-critical approach to competency management.

In this paper we present a pattern-based approach to the extraction of competencies in a knowledge-based research organization (scientific topics, technologies) from publicly available scientific publications. The core assumption of our approach is that such topics will not occur in random fashion across documents, but instead occur only in specific scientific discourse contexts that can be precisely defined and used as patterns for topic extraction.

The remainder of the paper is structured as follows. In section 2 we describe related work in competency management and argue for an approach based on natural language processing and ontology modeling. We describe our specific approach to topic extraction for competency management in detail in section 3. The paper then continues with the description of an experiment that we performed on topic extraction for competency management in our own organization, DFKI. Finally, we conclude the paper with some conclusions that can be drawn from our research and ideas for future work that arise from these.

2 Related Work

Competency management is a growing area of knowledge management that is concerned with the "identification of skills, knowledge, behaviors, and capabilities needed to meet current and future personnel selection needs, in alignment with the differentiations in strategies and organizational priorities." [1] Our particular focus here is on aspects of competency management relating to the identification and management of knowledge about scientific topics and technologies, which is at the basis of competency management.

Most of the work on competency management has been focused on the development of methods for the identification, modeling, and analysis of skills and skills gaps and on training solutions to help remedy the latter. An important initial step in this process is the identification of skills and knowledge of interest, which is mostly done through interviews, surveys and manual analysis of existing competency models. Recently, ontology-based approaches have been proposed that aim at modeling the domain model of particular organization types (e.g. computer science, health-care) through formal ontologies, over which matchmaking services can be defined for bringing together skills and organization requirements (e.g. [2], [3]).

The development of formal ontologies for competency management is important, but there is an obvious need for automated methods in the construction and dynamic maintenance of such ontologies. Although some work has been done on developing automated methods for competency management through text and web mining (e.g. [4]) this is mostly restricted to the extraction of associative networks between people according to documents or other data they are associated with. Instead, for the purpose of automated and dynamic support of competency management a richer analysis of competencies and semantic relations between them is needed, as can be extracted from text through natural language processing.

3 Approach

Our approach towards the automatic construction and dynamic maintenance of ontologies for competency management is based on the extraction of relevant competencies and semantic relations between them through a combination of linguistic patterns, statistical methods as used in information retrieval and machine learning and background knowledge if available.

Central to the approach as discussed in this paper is the use of domain-specific linguistic patterns for the extraction of potentially relevant competencies, such as scientific topics and technologies, from publicly available scientific publications. In this text type, topics and technologies will occur in the context of cue phrases such 'developed a tool for XY' or 'worked on methods for YZ', where XY, YZ are possibly relevant competencies that the authors of the scientific publication is or has been working on. Consider for instance the following excerpts from three scientific articles in chemistry:

…profile refinement method for nuclear and magnetic structures…
…continuum method for modeling surface tension…
…a screening method for the crystallization of macromolecules…

In all three cases a method is discussed for addressing a particular problem that can be interpreted as a competency topic: '*nuclear and magnetic structures*', '*modeling surface tension*', '*crystallization of macromolecules*'. The pattern that we can thus establish from these examples is as follows:

method for [TOPIC]

as in:

method for [nuclear and magnetic structures]
method for [modeling surface tension]
method for [(the) crystallization of macromolecules]

Other patterns that we manually identified in this way are:

approach for [TOPIC]
approaches for [TOPIC]
approach to [TOPIC]
approaches to [TOPIC]
methods for [TOPIC]
solutions for [TOPIC]
tools for [TOPIC]

We call these the 'context patterns', which as their name suggests provide the lexical context for the topic extraction. The topics themselves can be described by so-called 'topic patterns', which describe the linguistic structure of possibly relevant topics that can be found in the right context of the defined context patterns. Topic patterns are defined in terms of part-of-speech tags that indicate if a word is for instance a noun, verb, etc. For now, we define only one topic pattern that defines a topic as a noun (optional) followed by a sequence of zero or more adjectives followed by a sequence of one or more nouns. Using the part-of-speech tag set for English of the Penn Treebank [5], this can be defined formally as follows - JJ indicates an adjective, NN a noun, NNS a plural noun:

(.?)((NN(S)? |JJ)*NN(S)?)*

The objective of our approach is to automatically identify the most relevant topics for a given researcher in the organization under consideration. To this end we download all papers by this researcher through Google Scholar run the context patterns over these papers and extract a window of 10 words to the right of each matching occurrence.

We call these extracted text segments the 'topic text', which may or may not contain a potentially relevant topic. To establish this, we first apply a part-of-speech tagger (TnT: [6]) to each text segment and sub-sequentially run the defined topic pattern over the output of this. Consider for instance the following examples of context pattern, extracted topic text in its right context, part-of-speech tagged version[1] and matched topic pattern (highlighted):

approach to

semantic	*tagging*	*,*	*using*	*various*	*corpora*	*to*	*derive*	*relevant*	*underspecified*	*lexical*
JJ	***NN***	*,*	*VBG*	*JJ*	*NN*	*TO*	*VB*	*JJ*	*JJ*	*JJ*

semantic tagging

solutions for

anaphoric	*expressions*	*.*	*Accordingly*	*,*	*the*	*system*	*consists*	*of*	*three*	*major*	*modules*	*:*
JJ	***NNS***	*.*	*RB*	*,*	*DT*	*NN*	*VBZ*	*IN*	*CD*	*JJ*	*NNS*	*:*

anaphoric expressions

tools for

ontology	*adaptation*	*and*	*for*	*mapping*	*different*	*ontologies*	*should*	*be*	*an*
NN	***NN***	*CC*	*IN*	*VBG*	*JJ*	*NNS*	*MD*	*VB*	*DT*

ontology adaptation

approach for

modeling	*similarity*	*measures*	*which*	*tries*	*to*	*avoid*	*the*	*mentioned*	*problems*
JJ	***NN***	***NNS***	*WDT*	*VBZ*	*TO*	*VB*	*DT*	*VBN*	*NNS*

modelling similarity measures

methods for

domain	*specific*	*semantic*	*lexicon*	*construction*	*that*	*builds*	*on*	*the*	*reuse*
NN	***JJ***	***JJ***	***NN***	***NN***	*WDT*	*VBZ*	*IN*	*DT*	*NN*

domain specific semantic lexicon construction

[1] Clarification of the part-of-speech tags used: CC: conjunction; DT, WDT: determiner; IN: preposition; MD: modal verb; RB: adverb; TO: to; VB, VBG, VBP, VBN, VBZ: verb

As can be observed from the examples above, mostly the topic to be extracted will be found directly at the beginning of the topic text. However, in some cases the topic will be found only later on in the topic text, e.g. in the following examples[2]:

approach to
be used in a lexical choice system , the model of
VB VBN IN DT JJ NN NN , DT NN IN
lexical choice system

approach for
introducing business process-oriented knowledge management , starting on the …
VBG ***NN JJ NN NN*** *, VBG IN DT …*
business process-oriented knowledge management

The topics that can be extracted in this way now need to be assigned a measure of relevance, for which we use the well-known TF/IDF score that is used in information retrieval to assign a weight to each index term relative to each document in the retrieval data set [7]. For our purposes we apply the same mechanism, but instead of assigning index terms to documents we assign extracted topics (i.e. 'terms') to individual researchers (i.e. 'documents') for which we downloaded and processed scientific publications. The TF/IDF measure we use for this is defined as follows:

$$D = \{d_1, d_2, \ldots, d_n\}$$

$$D_{freq>1}^{topic} = \{d_1, d_2, \ldots, d_n\} \text{where } \mathrm{freq}_{\mathrm{d_i}}^{topic} > 1 \text{ for } 1 \le i \le n$$

$$tf_d^{topic} = \frac{freq_d^{topic}}{freq_D^{topic}}$$

$$idf^{topic} = \frac{|D|}{\left|D_{freq>1}^{topic}\right|}$$

$$tfidf_d^{topic} = tf_d^{topic} * idf^{topic}$$

where D is a set of researchers and $freq_{\mathrm{d}}^{topic}$ is the frequency of the topic for researcher d

The outcome of the whole process, after extraction and relevance scoring, is a ranked list of zero or more topics for each researcher for which we have access to publicly available scientific publications through Google Scholar.

[2] Observe that '*lexical choice system*' is a topic of relevance to NLP in natural language generation.

4 Experiment

To evaluate our methods we developed an experiment based on the methods discussed in the previous section, involving researchers from our own organization, DFKI. For all of these, we downloaded their scientific publications, extracted and ranked topics as explained above and then asked a randomly selected subset of this group to evaluate the topics assigned to them. Details of the data set used, the evaluation procedure, results obtained and discussion of results and evaluation procedure are provided in the following.

4.1 Data Set

The data set we used in this experiment consists of 3253 downloaded scientific publications for 199 researchers at DFKI. The scientific content of these publications are all concerned with computer science in general, but still varies significantly as we include researchers from all departments at DFKI[3] with a range of scientific work in natural language processing, information retrieval, knowledge management, business informatics, image processing, robotics, agent systems, etc.

The documents were downloaded by use of the Google API, in HTML format as provided by Google Scholar. The HTML content is generated automatically by Google from PDF, Postscript or other formats, which unfortunately contains a fair number of errors - among others the contraction of ‘fi’ in words like ‘specification’ (resulting in ‘specication’ instead), the contraction of separate words into nonsensical compositions such as ‘stemmainlyfromtwo’ and the appearance of strange character combinations such ‘â€“’. Although such errors potentially introduce noise into the extraction we assume that the statistical relevance assignment will largely normalize this as such errors do not occur in any systematic way. Needless to say that this situation is however not ideal and that we are looking for ways to improve this aspect of the extraction process.

The document collection was used to extract topics as discussed above, which resulted first in the extraction of 7946 topic text segments by running the context patterns over the text sections of the HTML documents[4]. The extracted topic text segments (each up to 10 words long) were then part-of-speech tagged with TnT, after which we applied the defined topic pattern to extract one topic from each topic text[5]. Finally, to compute the weight of each topic for each researcher (a topic can be assigned to several researchers but potentially with different weights) and to assign a

[3] See http://www.dfki.de/web/welcome?set_language=en&cl=en for an overview of DFKI departments and the corresponding range in scientific topics addressed.

[4] For this purpose we stripped of HTML tags and removed page numbering, new-lines and dashes at end-of-line (to normalize for instance ‘as-signed’ to ‘assigned’).

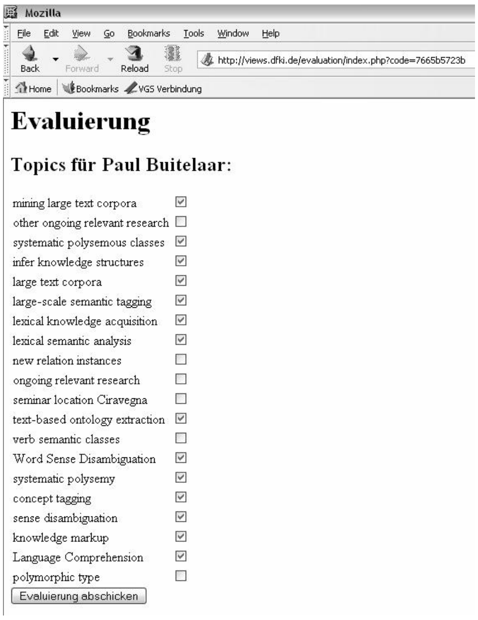
[5] In theory it could also occur that no topic can be identified in a topic text, but this will almost never occur as the topic text will contain at least one noun (that matches the topic pattern as defined in section 3).

ranked list of topics to each researcher, we applied the relevance measure as discussed above to the set of extracted topics and researchers.

4.2 Evaluation and Results

Given the obtained ranked list of extracted topics, we were interested to know how accurate it was in describing the research interests of the researchers in question. We therefore randomly selected a subset of researchers from the 199 in total that we extracted topics for, including potentially also a number of researchers without assigned topics, e.g. due to sparse data in their case. This subset of researchers that we asked to evaluate their automatically extracted and assigned topics consisted of 85 researchers, out of which 48 submitted evaluation results.

The evaluation consisted of a generated list of extracted and ranked topics, for which the researcher in question was asked simply to accept or decline each of the topics. The evaluation process was completely web-based, using a web form as follows:


Mozilla
File Edit View Go Bookmarks Tools Window Help
Back
Forward
Reload
Stop
http://views.dfki.de/evaluation/index.php?code=7665b5723b
Home
Bookmarks
VGS Verbindung

Evaluierung

Topics für Paul Buitelaar:

mining large text corpora ☑
other ongoing relevant research ☐
systematic polysemous classes ☑
infer knowledge structures ☑
large text corpora ☑
large-scale semantic tagging ☑
lexical knowledge acquisition ☑
lexical semantic analysis ☑
new relation instances ☐
ongoing relevant research ☐
seminar location Ciravegna ☐
text-based ontology extraction ☑
verb semantic classes ☐
Word Sense Disambiguation ☑
systematic polysemy ☑
concept tagging ☑
sense disambiguation ☑
knowledge markup ☑
Language Comprehension ☑
polymorphic type ☐

Figure 1: Web-form for evaluation of extracted topics

The evaluation for the 48 researchers that responded covered 851 extracted topics, out of which 380 were accepted as appropriate (44.65%). The following table provides a more detailed overview of this by distinguishing groups of researchers according to a level of how they judged their assigned topics correct (‘Level of Correctness’).

Level of Correctness	Number of Researchers
0-10%	*7*
11-20%	*1*
21-30%	*3*
31-40%	*9*
41-50%	*6*
51-60%	*9*
61-70%	*10*
71-80%	*3*
81-100%	*0*
	48

Table 1: Evaluation results

4.3 Discussion

Results of the evaluation vary strongly between researchers: almost half of them judge their assigned topics as more than 50% correct and 13 judge them more than 60% correct – on the other hand, 7 researchers are very critical of the topics extracted fro them (less than 10% correct) and slightly more than half judge their assigned topics less than 50% correct.

Additionally, in discussing evaluation results with some of the researchers involved we learned that it was sometimes difficult for them to decide on the appropriateness of an extracted topic, mainly because a topic may be appropriate in principle but it is: i) too specific or too general; ii) slightly spelled wrong; iii) occurs in capitalized form as well as in small letters; iv) not entirely appropriate for the researcher in question. We also learned that researchers would like to rank (or rather re-rank) extracted topics, although we did not explicitly tell them they were ranked in any order.

In summary, we take the evaluation results as a good basis for further work on topic extraction for competency management, in which we will address a number of the smaller and bigger issues that we learned out of the evaluation.

5 Applications

The overall application of the work presented here is management of competencies in knowledge organizations such as research institutes like DFKI. As mentioned we will therefore make the extracted topics available as ontology and corresponding knowledge base, on which further services can be defined and implemented such as expert finding and matching. For this purpose we need to organize the extracted topics further by extracting relations between topics and thus indirectly between researchers or groups of researchers working on these topics. We took a first step in this direction by analyzing the co-occurrence of positively judged topics (380 in total) from our evaluation set in the documents that they were extracted from. This resulted in a ranked listed of pairs of topics co-occurring more or less frequently. The following table provides a sample of this (the top 15 co-occurring topics over the 1091 documents for the 48 researchers that responded to the evaluation task):

# of co-occurrences	Topic 1	Topic 2
1164	knowledge representation	knowledge base
796	information retrieval	knowledge base
676	question answering	knowledge base
528	question answering	information retrieval
524	knowledge representation	information retrieval
416	business process	business process modeling
416	knowledge representation	context information
384	information retrieval	context information
368	context information	knowledge base
364	information retrieval	sense disambiguation
360	business process	information retrieval
336	knowledge representation	question answering
336	linguistic processing	information retrieval
296	business process	knowledge base
292	knowledge markup	knowledge base

Table 2: Top-15 co-occurring topics

We can also visualize this as follows:

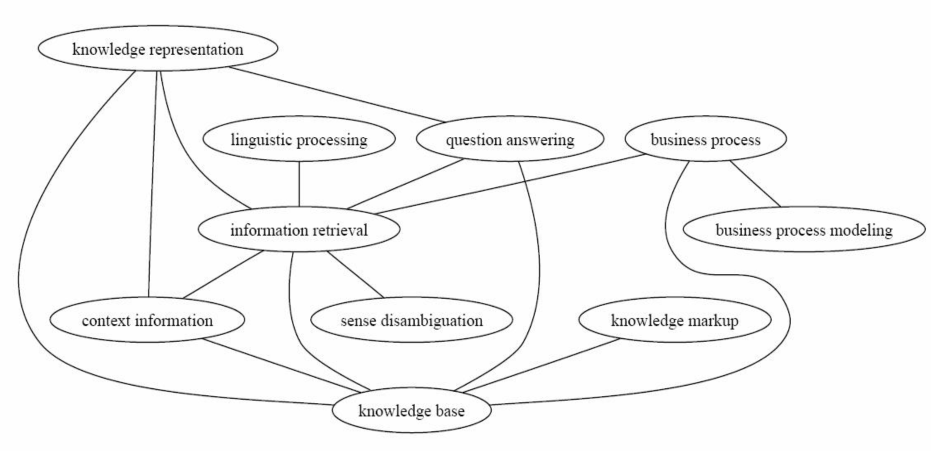
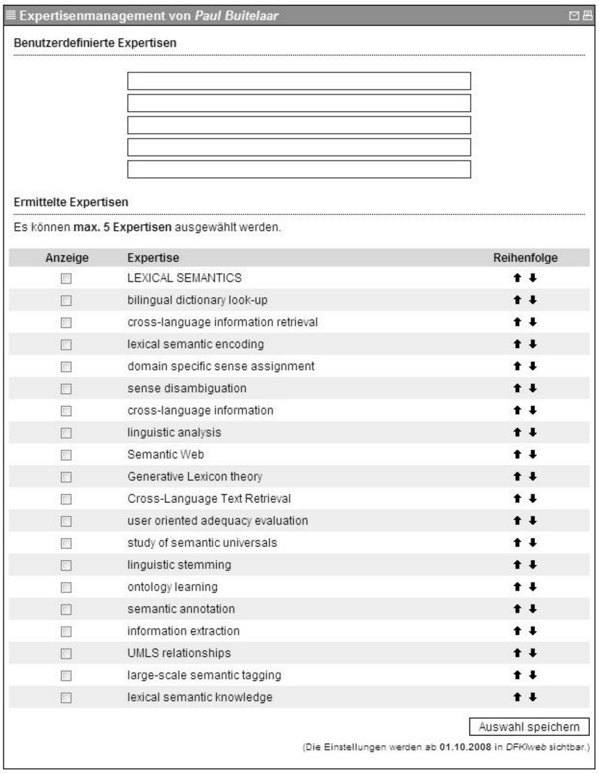


Figure 2: Association network between extracted topics (excerpt)

A different application that we are working on is to display the competencies of DFKI researchers in our web sites, e.g. by hyperlinking their names with an overview of competencies (scientific topics, technologies) that were either extracted automatically with the procedures discussed here or manually defined by the researchers themselves. For this purpose we integrate extracted topics into an individualized website on the DFKI intranet that allows each researcher to manage this as they see fit as follows:



Expertisenmanagement von *Paul Buitelaar*

Benutzerdefinierte Expertisen

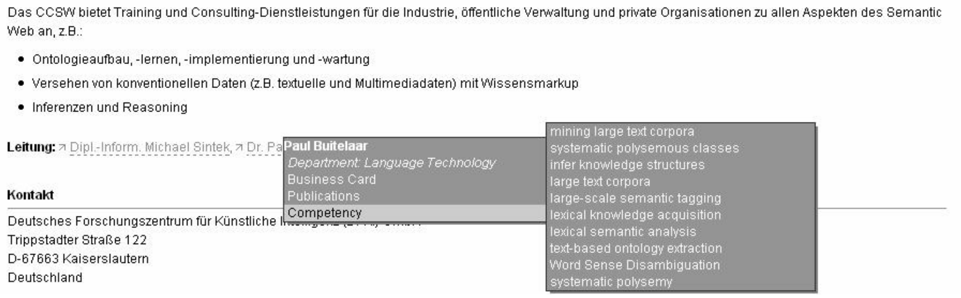
Ermittelte Expertisen

Es können **max. 5 Expertisen** ausgewählt werden.

Anzeige	Expertise	Reihenfolge
☐	LEXICAL SEMANTICS	↑ ↓
☐	bilingual dictionary look-up	↑ ↓
☐	cross-language information retrieval	↑ ↓
☐	lexical semantic encoding	↑ ↓
☐	domain specific sense assignment	↑ ↓
☐	sense disambiguation	↑ ↓
☐	cross-language information	↑ ↓
☐	linguistic analysis	↑ ↓
☐	Semantic Web	↑ ↓
☐	Generative Lexicon theory	↑ ↓
☐	Cross-Language Text Retrieval	↑ ↓
☐	user oriented adequacy evaluation	↑ ↓
☐	study of semantic universals	↑ ↓
☐	linguistic stemming	↑ ↓
☐	ontology learning	↑ ↓
☐	semantic annotation	↑ ↓
☐	information extraction	↑ ↓
☐	UMLS relationships	↑ ↓
☐	large-scale semantic tagging	↑ ↓
☐	lexical semantic knowledge	↑ ↓

(Die Einstellungen werden ab **01.10.2008** in *DFKIweb* sichtbar.)

Figure 3: DFKI Intranet web-form for personalized expertise management



Das CCSW bietet Training und Consulting-Dienstleistungen für die Industrie, öffentliche Verwaltung und private Organisationen zu allen Aspekten des Semantic Web an, z.B.:

- Ontologieaufbau, -lernen, -implementierung und -wartung
- Versehen von konventionellen Daten (z.B. textuelle und Multimediadaten) mit Wissensmarkup
- Inferenzen und Reasoning

Leitung: ↗ Dipl.-Inform. Michael Sintek, ↗ Dr. Pa

Kontakt

Deutsches Forschungszentrum für Künstliche I
Trippstadter Straße 122
D-67663 Kaiserslautern
Deutschland

mining large text corpora
systematic polysemous classes
infer knowledge structures
large text corpora
large-scale semantic tagging
lexical knowledge acquisition
lexical semantic analysis
text-based ontology extraction
Word Sense Disambiguation
systematic polysemy

Figure 4: DFKI Intranet web application for expertise visualization

6 Conclusions and Future Work

In this paper we described an approach towards automatic, dynamic and time-critical support for competency management based on topic extraction from relevant text documents. In the use case we presented, we focus on the extraction of topics that represent competencies in scientific research and technology. Results obtained through an experiment on this for our own organization, DFKI, as example of a knowledge organization, are encouraging and provided us with useful feedback for improving our methods further. In current and future work we are therefore addressing some of the issues encountered during the evaluation process, in particular on improving the quality of the document collection, extending the coverage and precision of the topic and context patterns and further experimenting with the ranking scores we use.

Besides this we are currently extending the work on relation extraction between topics and (groups of) researchers as presented in an early stage in section 5, leading to methods for exporting extracted topics and relations as a shallow ontology with a corresponding knowledge base of associated researchers and documents that can be used to build further services such as semantic-level expert finding and matching.

Finally, we are currently preparing an extended evaluation that will include comparison with a baseline method on topic extraction, which does not use any specific context patterns as we defined and used them in our approach. For this purpose we are considering the use of TermExtractor[6], which enables the extraction of domain-relevant terms from a corresponding domain-specific document collection [8]. We consider the task of term extraction vs. topic extraction to be similar enough to justify this comparison.

Acknowledgements

This work was supported by the DFKI-internal project VIeWs4DFKI and in part by the Theseus-MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under grant number 01MQ07016.

We would like to thank our colleagues at DFKI for providing feedback on the evaluation of automatically extracted topics and two anonymous reviewers for providing constructive and very useful feedback for improving this paper as well as our methods and evaluation in general.

References

1. Draganidis, F. Mentzas, G.: Competency based management: A review of systems and approaches. Information Management and Computer Security 14(1), pp. 51–64 (2006)

[6] http://lcl2.uniroma1.it/termextractor/

2. Colucci, S., Di Noia, T., Di Sciascio, E., Donini, F., Piscitelli, G., Coppi, S.: Knowledge based approach to semantic composition of teams in an organization. In: SAC-05, ACM, New York, pp. 1314–1319 (2005)
3. Kunzmann, C., Schmidt, A.: Ontology-based Competence Management for Healthcare Training Planning: A Case Study. In: I-KNOW 2006, Graz & Special Issue of the Journal of Universal Computer Science (J.UCS), ISSN 0948-695X, pp. 143-150 (2006)
4. Zhu, J., Goncalves, A. L., Uren, V. S., Motta, E., Pacheco, R.: Mining Web Data for Competency Management. In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), pp. 94-100 (2005)
5. Marcus, M., Santorini, B., Marcinkiewicz, M.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19(2) (1993)
6. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: 6th ANLP Conference, Seattle, WA (2000)
7. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management 24/5, pp.515-523 (1988)
8. Sclano, F., Velardi, P.: TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. In: 3rd International Conference on Interoperability for Enterprise Software and Applications I-ESA 2007, Funchal, Madeira Island, Portugal (2007)