Uwe Mönnich, Kai-Uwe Kühnberger

Editorial

1 Introduction

The division of scientific disciplines into a theory part and a more applied or practical part is a rather common conceptualization of classifying academic fields and can be found in nearly all academic research traditions ranging from the sciences and engineering disciplines to the humanities and educational research. Although there seems to be a natural order of producing scientific results, namely an order of how to develop theory and practice – first, a theory should be developed, i.e. a conceptualization of a certain domain must be provided, and second, this theory can be tested in experiments, implementations, applications etc. – there are many examples in the history of academic disciplines where theory follows practical developments and not vice versa. Text technology is perhaps such an example: Markup standards such as RDF, OWL, or XML, coding initiatives like OLAC (Open Language Archives Community), and practical applications for retrieval purposes (often in business–related contexts) seem sometimes to get ahead of theoretical characterizations of the underlying standards. For example, a standard like OWL Full is at present theoretically not very well understood and it took some time to specify the theoretical machine models of markup languages (like XML) – actually, at a time point after the languages themselves have been accepted as de facto standards.

Nevertheless, we decided to follow the natural order of categorizing cutting–edge research in text technology into theory and practice for this present double volume "Ontologies in Text Technology" of the LDV–Forum: In the first volume, theory–related papers are collected, whereas work shedding light on applications is covered in the present second volume. Although text technology itself and, in particular, its connection to coding semantic knowledge in form of ontologies is a rather young discipline and some practical developments seem to hurry ahead of their theoretical foundations, we think that this order enables the reader to follow a more logical succession of the recent developments. This is further supported by the fact that articles contained in the first volume can be considered in many aspects as a basis for results provided in this second volume. More will be said about these interrelationships between the two types of articles in Section 3.

In any case, we as the guest editors are proud to present the second part entitled "Applications of Ontologies in Text Technology" of the double issue of the LDV–Forum to the research community. We hope that the interested reader can profit from the work collected here and in the best of all possible worlds can pick up some ideas for her own research, in order to further promote text technology and ontology design.

2 The Research Unit 437 Text Technological Information Modeling

During the last six years the development of text technology in Germany was strongly influenced by the research unit 437 "Text Technological Information Modeling" funded by the German Research Foundation (DFG). This research unit is an interdisciplinary research endeavor carried by the Universities of Bielefeld, Gießen, Dortmund, Tübingen, and Osnabrück. Starting in the year 2001, this group constitutes the largest collaborative research project devoted to text technological issues and has provided the basis for text technological research in Germany. Currently this research unit is in its final funding year. In order to get a better idea of the overall project, a concise overview of the subprojects funded during the second phase of this research unit is given:

- *Secondary Structuring of Information and Comparative Analysis of Discourse.*
  Principal Investigator: Dieter Metzing.
- *Induction of Document Grammars for the Representation of Logical Hypertextual Document Structures.*
  Principal Investigator: Alexander Mehler.
- *Text-Grammatical Foundations for the (Semi-)Automated Text-to-Hypertext Conversion.*
  Principal Investigator: Angelika Storrer.
- *Generic Document Structures in Linearly Organized Texts: Text Parsing Using Domain Ontologies and Text Structure Ontologies.*
  Principal Investigator: Henning Lobin.
- *Adaptive Ontologies on Extreme Markup Structures.*
  Principal Investigators: Uwe Mönnich, Kai-Uwe Kühnberger.

Although the research unit tries to cover all aspects of current text technological activities, it is still possible to identify certain core aspects that play a central role in all subprojects. Examples for such vertical topics of the whole research unit are ontologies, annotations, markup standards, and processing aspects of texts. All these topics play an important role in all participating projects. Some aspects of these vertical topics of the research unit are also represented in this double volume of the LDV-Forum. Whereas certain sub-projects of the collaborative research unit mentioned above are represented in Volume I focusing on the foundations of theories for developing, characterizing, coding, learning, and adapting ontological background knowledge as a crucial challenge for the semantic annotation of text documents, other sub-projects document aspects of their ongoing work in the present Volume II "Applications of Ontologies in Text Technology". We think that we can provide in this way not only a representative documentation of text technology in general, but also a representative collection illustrating the research unit 437, in particular.

3 The Structure of Volume II

This second volume collects applied work on ontology design and text technology. The rticles span a field from ontologies in discourse parsing and lexical semantics to anaphora resolution, linguistic annotations, and the automatic acquisition of formal concepts

from textual data. It is important to notice that there are many connections between the articles published in the two parts of the double volume. In particular, several foundational results presented in the first volume provided the basis for applications in the present volume. We will try to make some of these obvious connections visible while roughly summarizing important topics of the contributions collected here.

In his article "An Ontology of Linguistic Annotations", Christian Chiarcos discusses necessary design features of ontological resources for annotations mainly intended for terminological integration, and ontology–based search across linguistic resources with heterogeneous annotations. By developing a structured ontology involving selfcontained sub-ontologies, which are linked in a declarative way, he shows how a separation between the annotation documentation and its interpretation with respect to the reference terminology can be achieved. The underlying idea is a mapping process of annotations onto ontological representations, such that the full range of types of information in annotations (like syntactic, semantic, phonological etc. information) can be referenced by an ontology. A theoretical basis of the ideas spelled out in Chiarcos article can be found in the contribution of the first volume "Towards a Logical Description of Trees in Annotation Graphs" by Jens Michaelis and Uwe Mönnich.

The contribution "OWL Ontologies as a Resource for Discourse Parsing" by Maja Bärenfänger, Mirco Hilbert, Henning Lobin, and Harald Lüngen bases discourse annotations on the Rhetorical Structure Theory (Mann and Thompson, 1988) to automatically arrange discourse segments and rhetorical relations in a tree structure. The resources to extract these tree structures are based on heterogeneous types of information like discourse marker lexicons, lexico-semantic ontologies, and annotation layers of input text. The article focuses particularly on OWL ontologies and how they can be consulted by the discourse parser. An important role plays the usage of an OWL version of GermaNet and a taxonomy of rhetorical relations which was developed by the authors themselves. Certain aspects of this paper are practical applications of the theoretical results of the contribution "Domain Ontologies and Wordnets in OWL: Modelling Options" by Harald Lüngen and Angelika Storrer of the first volume.

The development of automatic extraction procedures for generating cheap but nevertheless reliable ontologies seems to be one of the most important practical challenges for text technological research (Perez and Mancho, 2003). In particular, synonymy information seems to be a good starting point for such an endeavor to identify different candidates for one and the same concept (word sense). A Kumaran, Ranbeer Makin, Vijay Pattisapu, Shaik Sharif, and Lucy Vanderwende examine in their article "Evaluating the Quality of Automatically Extracted Synonymy Information" two complementary techniques in order to automatically extract synonymy information from large corpora: First, a generic broad–coverage parser for generating bits of semantic information and second, their synthesis into sets of synonyms using word-sense disambiguation with latent semantic analysis. The authors evaluate their approaches quantitatively and qualitatively. From a general perspective this article is a further step towards the whole cycle of automatic ontology generation: extracting semantic information, expanding an ontology with additional information, and adapting this expanded ontology if necessary. In this

sense, the present paper complements the article "Automatic Ontology Extension: Resolving Inconsistencies" by Ekaterina Ovchinnikova and Kai-Uwe Kühnberger in the first volume.

A classical challenge of natural language processing concerns nominal anaphora resolution, partially because several types of knowledge have to be taken into account as, for example, morphosyntactic information and domain knowledge. The paper "Resolving Nominal Anaphora Using Hybrid Semantic Knowledge" by Daniela Göcke, Maik Stührenberg, and Tonio Wandmacher proposes a hybrid approach towards extracting automatically necessary domain knowledge: first, they propose a knowledge-free approach of distributional similarity (Paaß et al., 2004) based on latent semantic analysis, in this respect comparable to the paper "Evaluating the Quality of Automatically Extracted Synonymy Information" above, and second, they use Hearst patterns (Hearst, 1982), i.e. predicate-argument relations encoded in the syntactic structure of the text. The integration of semantic relatedness by combining information about extracted relations and cooccurrence information is used to identify the most likely antecedent in anaphora resolution tasks. The authors evaluate their approach on a corpus of German scientific and newspaper articles.

The final contribution of the present volume "Automatic Acquisition of Formal Concepts from Text" by Pablo Gamallo Otero, Gabriel Pereira Lopes, and Alexandre Agustini uses formal concept analysis (Priss, 2006) in order to implement an unsupervised learning procedure for concept acquisition from annotated corpora. The idea is to build bidimensional clusters of words and their lexico–semantic contexts. Their procedure results in a concept lattice describing a domain–specific ontology underlying the training corpus. The authors use for their evaluation a large Portuguese corpus where the tokens were extracted from a general–purpose journal and an English excerpt of the European Parliament Proceedings.

4 Acknowledgments

This volume would not have been possible without the help of many people. In the first place, the guest editors want to thank the editors-in-chief of the LDV–Forum, Alexander Mehler and Christian Wolff. Their encouragement and support in all phases of the emergence of this double volume has been irreplaceable in completing it. Furthermore we want to thank the German Research Foundation for financial support of the research unit 437 "Text Technological Information Modeling" and particularly the speaker of this research unit, Dieter Metzing.

Last but not least, the editors want to thank the program committee for their careful evaluations of the submitted papers. The quality of this volume is also a direct reflection of the work these reviewers invested. The program committee consisted of the following researchers (in alphabetical order): Irene Cramer, Thierry Declerck, Stefan Evert, Pascal Hitzler, Wolfgang Höppner, Helmar Gust, Marcus Kracht, Edda Leopold, Alessandro Moschitti, Larry Moss, Rainer Osswald, Olga Pustylnikov, Georg Rehm, Hans-Christian Schmitz, Bernhard Schröder, Uta Seewald-Heeg, Manfred Stede, Markus Stuptner, Frank Teuteberg, Yannick Versley, Johanna Völker, Armin Wegner, and Christian Wolff.

References

Hearst, M. (1982). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics.

Mann, W. and Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organsation. Text, 8(3):243–281.

Paaß, G., Kindermann, J., and Leopold, E. (2004). Learning prototype ontologies by hierarchical latent semantic analysis. In Knowledge Discovery and Ontologies, Pisa, Italy.

Perez, G. A. and Mancho, M. D. (2003). A Survey of Ontology Learning Methods and Techniques. OntoWeb Delieverable 1.5.

Priss, U. (2006). Formal concept analysis in information science. Information Science and Technology, 40:521–543.