Knowledge-Based Support for Scholarly Editing and Text Processing

Kittelmann, Jana
info@janakittelmann.de
MLU Halle-Wittenberg, Deutschland

Wernhard, Christoph
info@christophwernhard.com
TU Dresden, Deutschland

Introduction

Background: Large Knowledge Bases

A large portion of the material on which scholarly editing is based today is available electronically in large knowledge bases. Some of these emerge from the archive, library and museum communities, for example *Kalliope*. Such efforts require the use of standardized vocabularies and databases of entities such as persons and locations. *Kalliope* thus links to *Gemeinsame Normdatei (GND)*, which provides more than 120 million facts about approximately 11 million entities. The prevailing technique to realize such linked knowledge bases is the Semantic Web, as advocated by the W3C, characterized by the use of ontologies to express standardized vocabularies, global identifiers (URIs) and the possibility to express knowledge in a machine understandable way as subject-predicate-object statements with RDF. Further large knowledge bases, such as *Yago* (Hoffart et al. 2013) and *DBpedia* (Lehmann et al. 2015), developed mainly in computer science with Semantic Web techniques, gather and combine machine processable knowledge from "crowd-maintained" sources like *Wikipedia* and centrally maintained sources like *GND* or *GeoNames*.

Beyond *TEI*

The seemingly best developed machine support for scholarly editing today is provided with the *Text Encoding Initiative (TEI)* format, based on document markup. URIs as attribute values of markup elements can provide links to knowledge bases. Envisaged applications include in particular the rendering for different media and extraction of metadata. Some of the recent developments are actually orthogonal to the OCHCO text model and its representation through XML, core characteristics of the original *TEI*. Connecting *TEI* with Semantic Web techniques, data modeling and ontologies is, for example, an ongoing topic of discussion (e.g. Eide 2015). Recent versions of *TEI* provide support for *names, dates, people, and places* as well as *linking, segmentation, and alignment* (The TEI Consortium 2015: Chapters 13 and 16). In a broad long-term perspective, important aspects that further go into these directions become apparent:

- Incorporation of advanced semantics related techniques such as named entity recognition or statistics-based text analysis.
- Relationships to external knowledge bases and to formal semantics.
- Obtaining high-quality presentations without requiring expensive development of dedicated XML transformations and stylesheets.
- Loose coupling of object text and markup: Alternate markup by different authors or for different purposes should be supported. Markup generated by automated methods should not clutter up the document. Queries and transformations should remain applicable also after changes of the markup. Sustainability must not be compromised by dependency on short-lived technology and specifications.

Addressing these issues, we approach the requirements of today's scholarly editing here from the view of computational logic: What can logics – as machine processable symbolic languages with formally specified semantics – contribute? A starting point is that with Semantic Web technology the large knowledge bases can already be considered as large sets of logic facts. Logic languages have various further potential roles in machine supported scholarly editing, such as specifying properties and values associated with texts, specifying pieces of text, specifying knowledge sources and their combination, and specifying inferences involved in automated computation of information associated with texts.

Knowledge-Based Support for Scholarly Editing

High-Quality Support at all Phases

Three main phases of machine assisted scholarly editing can be identified, which all should be supported: (1) Creating the enhanced object text; (2) Generating intermediate representations for inspection by humans or machines; (3) Generating consumable presentations. Support for all three phases should be of high quality – for example entity recognition should precisely identify persons, or the print layout of a finally rendered document should be professional.

Issues of Integrating Different Types of Knowledge

High-quality support is not possible without inclusion of specialized techniques and the combination of automated techniques with information and adjustments provided by humans. The adequate support of this combination is an important aspect where the considered scenario differs from conventional programming or query languages. Relevant techniques include non-monotonic reasoning, semantics-based knowledge partitioning (Wernhard 2004, Ghilardi et al. 2006, Cuenca Grau et al. 2008, Kontchakov et al. 2010) and the use of explanations for inferred information, as exemplified by proofs in mathematical knowledge bases (Urban et al. 2013). A further important integration requirement concerns the combination of statistics-based techniques, which are essential for natural language processing operations such as named entity recognition or keyphrase extraction, with a symbolic logic-based framework.

External Annotations

The availability of powerful techniques to identify places in text – based on syntactic as well as semantic properties – suggests to prefer external annotations to in-place markup. Annotations are then maintained separated from the object text in annotation documents. An automated processor creates an annotated document by merging annotations and object text.

Representation of Epistemic Status

Scholarly editing requires to associate various forms of epistemic status with facts, which is interesting to model formally from the viewpoint of artificial intelligence. Consider for example a creation date associated with written communication: it can be given by its author or can be inferred – by the editor or by a machine, it can be only partially specified by the author, it can be specified with different precision, considered as a point or range in time, etc. The current version of *TEI* offers some related elements to indicate certainty, precision and responsibility (The TEI Consortium 2015: Chapter 21), but these are not based on any formal semantic treatment and it is seems hardly possible to express the sketched date examples with them.

Utilizing Inferred Access Patterns

Efficient access to large knowledge bases requires caching and preprocessing, which ideally should be performed automatically on the basis of the queries performed by the knowledge processing engine. Relevant techniques come from optimization in databases (Toman / Weddell 2011) and in first-order model computation systems (Pelzer / Wernhard 2007). It seems that recent techniques for view-based query processing (Calvanese et al. 2007) based on variants of Craig's interpolation and second-order quantifier elimination (Toman / Weddell 2011; Bárány et al. 2013; Wernhard 2014) where access patterns can be specifically considered in an abstract way (Bárány et al. 2013) are particularly useful. Logic-based languages for programming as well as data access facilitate the application of such abstract techniques. For an overview on alternate ways to associate computational meaning with logics see (Kowalski 2014).

The Role of Ontologies

Ontologies are an important ingredient for the Semantic Web because they provide agreed vocabularies. However, to evaluate queries arising in the text processing tasks of scholarly editing, ontology reasoning alone is not sufficient. Also, the basic ontologies relevant in the context of scholarly editing are – in contrast to the biomedical area (Horrocks 2013) – rather small and trivial.

A Prototype: The *KBSET* System

Important issues of complex computer systems often become apparent only with applications. Thus, the authors developed the *KBSET* system, an experimental platform to clarify the precise requirements of machine support for scholarly editing and to experiment with advanced techniques. It follows the outlined approach, but, so far, only realizes some of the discussed aspects. A draft version of an edition of *Max Stirner: Geschichte der Reaction, Band 1. Berlin, 1852* accompanies it as comprehensive example. The system is free software and available from http://cs.christophwernhard.com/kbset/.

In a typical setting, the system takes as inputs:

- A source text file, possibly in *LaTeX* format. The system can parse *LaTeX*, where the set of recognized commands is configurable, including user defined commands as well as commands that establish some "ordered hierarchy of content objects". In this way plain or structured text is available within the system to modules that operate on such text models.
- *Annotation documents*, that is, text files with annotations, possibly in *LaTeX* format. The associated places in the source text to which they are referring are specified abstractly.
- Large fact bases, currently in particular *GND* and *GeoNames*, as well as extracts from *YAGO2* and *DBpedia*.
- A so-called *assistance document*, that is, a configuration file, where, among other things, the fact bases are specified and information is given to bias or override automated inferencing such that fully correct results are obtained.

A user interface is provided that integrates the system into the *Emacs* editor, which is free software. The system includes a facility for named entity recognition, which –

essentially based on *GND* and *GeoNames* as gazetteers – identifies persons, locations and dates. The system produces a variety of outputs, supporting all the phases of scholarly editing mentioned above:

- *LaTeX* documents where annotations and inferred information are merged in. By passing unrestricted *LaTeX* access to the user, high-quality layouts can be achieved.
- Support during development by possibilities to highlight and inspect entities recognized by the system.
- An export possibility to visualize detected locations mentioned in the source text with the *Dariah* geobrowser.

A typical application would be the development of an annotated essay or book, where the source text is edited in *LaTeX* and the configuration evolves step-by-step until the inferred information is fully correct.

Acknowledgments

This work was supported by *Alexander von Humboldt-Professur für neuzeitliche Schriftkultur und europäischen Wissenstransfer* and by *DFG grant WE 5641/1-1* .

Bibliographie

Bárány, Vince / Benedikt, Michael / ten Cate, Balder (2013): "Rewriting guarded negation queries", in: *Mathematical Foundations of Computer Science 2013 (MFCS 2013)*, volume 8087 of LNCS. Berlin / Heidelberg / New York: Springer 89-110.

Calvanese, Diego / De Giacomo, Giuseppe / Lenzerini, Maurizio / Vardi, Moshe Y. (2007): "View-based query processing: On the relationship between rewriting, answering and losslessness", in: *Theoretical Computer Science* 371, 3: 169-182.

Cuenca Grau, Bernardo / Horrocks, Ian / Kazakov, Yevgeny / Sattler, Ulrike (2008): "Modular reuse of ontologies: Theory and practice", in: *Journal of Artificial Intelligence Research* 31: 273-318.

Eide, Øyvind (2015): "Ontologies, data modeling, and TEI", in: *Journal of the Text Encoding Initiative* 8.

Ghilardi, Silvio / Lutz, Carsten / Wolter, Frank (2006): "Did I damage my ontology? A case for conservative extensions in description logics", in: Doherty, Patrick / Mylopoulos, John / Welty, Christopher A. (eds.): *Proc. 10th Int. Conf. on Principles of Knowledge Representation (KR'06)*. Cambridge, MA: AAAI Press 187-197.

Hoffart, Johannes / Suchanek, Fabian M. / Berberich, Klaus / Weikum, Gerhard (2013): "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia", in: *Artificial Intelligence* 194: 28-61.

Horrocks, Ian (2013): "What are ontologies good for?", in: Kuppers, Bernd Olaf / Hahn, Udo / Artmann, Stefan (eds.): *Evolution of Semantic Systems*. Berlin / Heidelberg / New York: Springer 175-188.

Kontchakov, Roman / Wolter, Frank / Zakharyaschev, Michael (2010): "Logic-based ontology comparison and module extraction, with an application to DL-Lite", in: *Artificial Intelligence* 174, 15: 1093-1141.

Kowalski, Robert A. (2014): "Logic Programming", in: Siekmann, Jörg (ed.): *Computational Logic* (= Handbook of the History of Logic 9). Amsterdam: Elsevier 523-569.

Lehmann, Jens / Isele, Robert / Jakob, Max / Jentzsch, Anja / Kontokostas, Dimitris / Mendes N., Pablo / Hellmann, Sebastian / Morsey, Mohamed / van Kleef, Patrick / Auer, Sören / Bizer, Christian (2015): "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia", in: *Semantic Web* 6, 2: 167-195.

Pelzer, Björn / Wernhard, Christoph (2007): "System description: E-KRHyper", in: *Automated Deduction* (CADE-21), volume 4603 of LNCS (LNAI). Berlin / Heidelberg / New York: Springer 503-513.

The TEI Consortium (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 2.8.0* TEI Consortium http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 9. Oktober 2015].

Toman, David / Weddell, Grant (2011): *Fundamentals of Physical Design and Query Compilation San Rafael*. CA: Morgan and Claypool.

Urban, Josef / Rudnicki, Piotr / Sutcliffe, Geoff (2013): "ATP and presentation service for Mizar formalizations", in: *Journal of Automated Reasoning* 50 (2): 229-241.

Wernhard, Christoph (2004): "Semantic knowledge partitioning", in: *Logics in Artificial Intelligence*: 9th European Conf. (JELIA 04), volume 3229 of LNCS (LNAI). Berlin / Heidelberg / New York: Springer 552-564.

Wernhard, Christoph (2014): *Expressing view-based query processing and related approaches with second-order operators*", Technical Report - Knowledge Representation and Reasoning 14-02, TU Dresden, http://www.wv.inf.tu-dresden.de/Publications/2014/report-2014-02.pdf [letzter Zugriff 9. Oktober 2015].