Representational Interoperability of Linguistic and Collaborative Knowledge Bases

Konstantina Garoufi and Torsten Zesch and Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Computer Science Department
Technische Universität Darmstadt
Hochschulstr. 10, D-64289 Darmstadt, Germany
{garoufi,zesch,gurevych}@tk.informatik.tu-darmstadt.de

Creating a Natural Language Processing (NLP) application often requires to access lexical-semantic Knowledge Bases (KBs). Recently, Collaborative Knowledge Bases (CKBs) such as Wikipedia and Wiktionary[1] have been recognized as promising lexical-semantic KBs for NLP (Zesch et al., 2008b), complementing traditional Linguistic Knowledge Bases (LKBs). As CKBs differ significantly from LKBs concerning their content, structure and topological properties, the interoperability between CKBs and LKBs has become a major issue.

To address this problem, we have developed a model of representational interoperability between LKBs and CKBs, which abstracts over the differences in their structures, and enables a uniform representation of their content in terms of *entities* and lexical-semantic *relations* between them. An *entity* consists of a set of *lexeme–sense* pairs along with a part-of-speech (*PoS*). The currently supported *relations* are the lexical relations synonymy and antonymy, as well as the semantic relations hypernymy, hyponymy, holonymy, meronymy and other, which covers any lexical-semantic relation other than the previously listed. NLP algorithms can thus be implemented in an one-time effort, as they only have to "know" about generalized *entities* and *relations* instead of being adapted to each KB individually.

The KBs currently integrated are the LKBs WordNet (Fellbaum, 1998), GermaNet (Kunze, 2004), Cyc (Lenat and Guha, 1989), Roget's Thesaurus (Jarmasz and Szpakowicz, 2003), Leipzig Annotation Project (Biemann, 2005), and the CKBs Wikipedia and Wiktionary, which are available for a large number of languages. Some of these KBs are rich in linguistic knowledge extending beyond the lexical-semantic level, even forming a complex ontology in the domain of human consensus reality (e.g. Cyc). Our work, however, aims at the representation of the lexical-semantic knowledge level of the KBs, and not at the complete modeling of their contents. Moreover, at the moment it addresses solely the issue of structural interoperability of KBs rather than attempting content mappings between them. For this reason the model is free of potential mapping errors, conflicts or loss of information.

The system architecture of a typical NLP application using the representational interoperability interface is presented in Figure 1. Each KB implements the generic representational interoperability interface[2] by means of its native application programming interface (API). As concepts and relations are modeled differently in each KB, they are mapped onto *entities* and *relations*. For example, a synset from the LKB WordNet is mapped to an *entity* by adding each synonym from the synset as a *lexeme* in the *entity*, together with its *sense* number and its *PoS*. Likewise, an article from the CKB Wikipedia is mapped to an *entity* by adding the article name and all redirects as *lexemes*. In this case, *sense* and *PoS* are left unspecified, as this information cannot be directly retrieved from Wikipedia. Similarly, the encoded relations between WordNet synsets or Wikipedia articles are mapped onto the given set of lexical and semantic *relations*. Additional information originally related to the concepts, e.g. glosses or examples, does not belong to our representation of an *entity*, but still remains programmatically accessible.

[1] http://www.{wikipedia,wiktionary}.org

[2] Implemented as a Java interface.

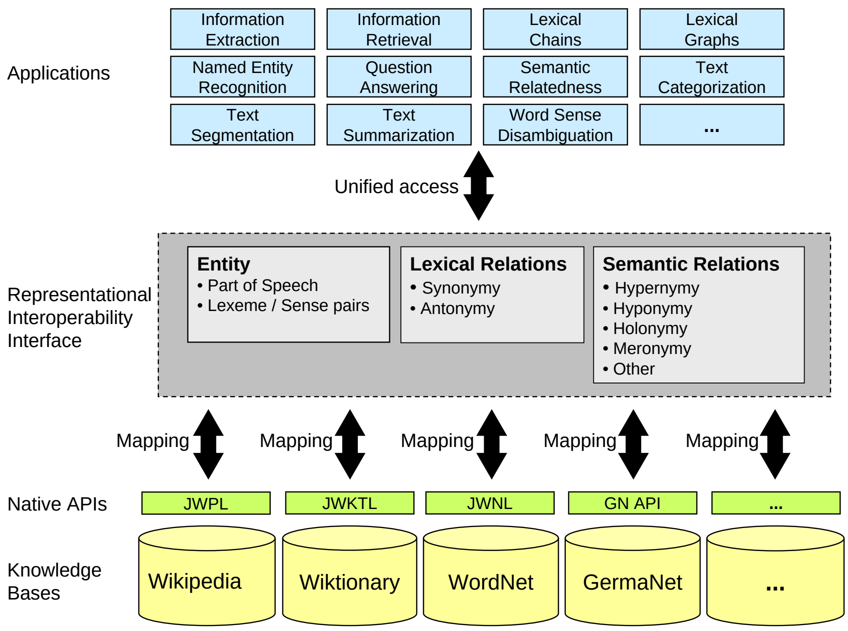

Figure 1: System architecture enabling representational interoperability.

We have so far employed the interoperability interface for *(i)* the computation of semantic relatedness (Zesch et al., 2008a), *(ii)* the construction of lexical chains and graphs (Schwager, 2008), and *(iii)* the graph-theoretic analysis of LKBs and CKBs (Garoufi et al., 2008). However, other applications relying on KBs can also benefit from it.

To our knowledge, there is no other framework of representational interoperability between LKBs and CKBs that has been designed from an application-oriented rather than a user-oriented perspective. The LEXUS tool for manipulating lexical resources (Kemps-Snijders et al., 2006), for example, which implements the common standardized Lexical Markup Framework (ISO TC37/SC4) for the construction of NLP lexicons (Francopoulo et al., 2006), is targeted at field linguists involved in language documentation rather than developers of NLP software. Other related work focuses on combining KBs on the content level in order to produce an enriched KB of greater coverage by merging or mapping concepts (Fröhner et al., 2005; Shi and Mihalcea, 2005; Suchanek et al., 2007; Medelyan and Legg, 2008). Our approach, in contrast, makes the first step toward a combination of a wide range of lexical-semantic KBs at a representational level, which supports practical NLP tasks, and can be extended to the content level in future work.

To conclude, we have presented a representational interoperability interface that implements a generalized model of lexical-semantic KBs, where the content of CKBs and LKBs is uniformly expressed in terms of *entities* and *relations*. Clearly, this generalized model cannot support the same level of expressiveness as directly accessing a KB. However, we believe that this is compensated for by the following advantages: *(i)* each NLP algorithm operating on a KB has to be implemented only once and can then be applied to all KBs, *(ii)* experimental results obtained using different KBs are better comparable, and *(iii)* the representational interoperability interface provides a framework for further work on full interoperability (including content alignment) of CKBs and LKBs.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) under the grants No. GU 798/3-1 and GU 798/1-3, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.

References

Chris Biemann. Semantic Indexing with Typed Terms Using Rapid Annotation. In *Proceedings of the TKE Workshop on Methods and Applications of Semantic Indexing*, Copenhagen, Denmark, August 2005.

Christiane Fellbaum. *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. Lexical Markup Framework (LMF). In *Proceedings of the Conference on Language Resources and Evaluation*, pages 233–236, Genoa, Italy, 2006.

Tina Fröhner, Matthias Nickles, Gerhard Weiß, Wilfried Brauer, and Rolf Franken. Integration of Ontologies and Knowledge from Distributed Autonomous Sources. *Künstliche Intelligenz*, pages 18–23, 2005.

Konstantina Garoufi, Torsten Zesch, and Iryna Gurevych. Graph-Theoretic Analysis of Collaborative Knowledge Bases in Natural Language Processing. In *Poster and Demo Session Proceedings of the 7th International Semantic Web Conference*, October 2008. (To appear).

Mario Jarmasz and Stan Szpakowicz. Roget's thesaurus and semantic similarity. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, pages 212–219, Borovets, Bulgaria, September 2003.

Marc Kemps-Snijders, Mark-Jan Nederhof, and Peter Wittenburg. Lexus, a web-based tool for manipulating lexical resources. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 1862–1865, Genoa, Italy, 2006.

Claudia Kunze. *Lexikalisch-semantische Wortnetze*, chapter Computerlinguistik und Sprachtechnologie, pages 423–431. Spektrum Akademischer Verlag, 2004.

Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

Olena Medelyan and Catherine Legg. Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, Chicago, USA, July 2008.

Florian Schwager. Automatic Analysis of Lexical Cohesion. Diploma thesis, Technische Universität Darmstadt, 2008.

Lei Shi and Rada Mihalcea. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 100–111, Mexico City, Mexico, February 2005.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th international World Wide Web conference*, Banff, Canada, May 2007.

Torsten Zesch, Christof Müller, and Iryna Gurevych. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the twenty-third AAAI Conference on Artificial Intelligence*, pages 861–867, Chicago, Illinois, July 2008a.

Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the sixth international Conference on Language Resources and Evaluation*, Marrakech, Morocco, May 2008b.