Book Review

Linked Lexical Knowledge Bases
Foundations and Applications

Iryna Gurevych, Judith Eckle-Kohler, and Michael Matuschek
(Technische Universität Darmstadt, Germany)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 34), 2016, xxi+124 pp; paperbound, ISBN 978-1-62705-974-9; ebook, ISBN 978-1-62705-904-6; doi 10.2200/S00717ED1V01Y201605HLT034, $50.00

Reviewed by
Maud Ehrmann
EPFL Digital Humanities Laboratory

Automatic text understanding requires knowledge and, so far, machines know only what we give or teach them. As a consequence, most natural language processing (NLP) tasks crucially rely on the existence of linguistic resources that encode information about language, be it of morphological, syntactic, or semantic nature. Such resources are typically acquired via two main approaches: the *knowledge-based* approach, or top–down, where information is manually curated by humans, and the *corpus-based* approach, or bottom–up, where information is automatically learned from corpora. Although the latter has gained ground during the last decade—benefiting from the availability of large amounts of text and from increased computing capacities—the former remains fundamental for it allows us to collect reliable, fine-grained, and explicit information.

Lexical knowledge bases (LKBs), also known as lexico-semantic resources, provide information about words and potentially entities, and are at the core of knowledge-based approaches. They are widely used in a variety of NLP tasks (e.g., word sense disambiguation, information retrieval, and question answering), all the more so since their traditional limitations (i.e., lack of language and domain-specific coverage) have recently started to fall. Indeed, beside the long-established process of expert-based resource creation (e.g., WordNet), Web technologies have enabled the collaborative, crowd-based construction of resources (e.g., Wikipedia and Wiktionary). This contributed to significantly widen the scope of the available machine-readable knowledge and, in the context of an already diverse landscape of LKBs, it encouraged and motivated even more the need to integrate different resources so as to make the best of them all.

This book introduces linked lexical knowledge bases by giving an account of their foundations and presenting their main applications. Its target audience includes NLP practitioners or students who wish to better understand linked lexical knowledge bases, how they are built, and their typical usages and added value. The book is organized into eight chapters plus a preface, and is additionally put into perspective by a foreword (by Ido Dagan) that considers the history, recent evolution, and probable future directions of knowledge acquisition.

doi:10.1162/COLI_r_00289

© 2017 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

Chapter 1 presents lexical knowledge bases at large. It first defines an LKB and reviews the different types of information it can hold. Then, main representatives of LKBs are introduced, considering first expert-built LKBs, followed by collaboratively constructed ones. A variety of LKBs are thus presented, from WordNets to FrameNets via Wiktionary and OmegaWiki, and, for each one, information types, main advantages, and typical usages are explained. Finally, a small section is dedicated to lexical knowledge representation standards; however, the authors specify that it is a rather side topic. Overall, besides giving the reader the necessary background for understanding the role and the importance of LKBs, this chapter makes the point of the great heterogeneity of information types and their organization in LKBs, thereby introducing the rationale for their joint use.

Chapter 2 introduces linked lexical knowledge bases (LLKB). It starts with the basic consideration of what does linking two LKBs mean, and at which level can it be done. As a consequence, two formal definitions are introduced, one of *word sense linking*, and the other of *linked lexical knowledge base*. Given these fundamentals, the chapter moves on with the consideration of previous efforts to combine LKBs. The earliest work mentioned dates back to 1994 with an attempt to link WordNet to the Longman Dictionary of Contemporary English; later on, linking between expert-built LKBs gradually gives way to the integration of collaboratively built resources. After this overview, the chapter focuses more specifically on two existing large-scale LLKBs, namely, *Uby* and *BabelNet*, and explains how they complement each other. Their different philosophies are indeed nicely grasped and presented with, among others, the fact that Uby was originally designed to cover verb information and models sources separately, whereas BabelNet initially focused on nouns, targeting a high level of multilinguality and combining information in a unified frame. Lastly, the chapter considers manual and collaborative linking of resources, as a transition towards their necessary automatic alignment.

Chapter 3 represents the core of the book and is dedicated to linking algorithms. To better specify what corresponds to the task of LKB linking, the chapter first presents what it is *not* and differentiates it from ontology matching, database schema matching, and graph matching. Once this framework is established, evaluation metrics and main approaches for automatic word sense alignment are introduced. The task of computing the similarity between two word senses usually relies on the exploitation of either sense definitions (gloss similarity-based word sense linking), resource's structure (graph structure-based word sense linking), or both (joint modeling). The feasibility and efficiency of an approach naturally depend on the type of information a resource contains, its degree of structuring, and the covered language(s). Each of these aspects is well explained and documented, with definitions and reviews of existing work.

Chapter 4 discusses the added value of LLKBs with respect to the widely defined task of "textual units disambiguation," which encompasses here word sense disambiguation, entity linking, and semantic role labeling. Fundamental disambiguation approaches are detailed (Lesk-based, graph-based, and machine learning-based) and an overview of work in each area is presented. It is shown that both rule-based and graph-based approaches benefit from the usage of LLKBs, for they provide richer sense representations with, for example, aggregated gloss information or dense networks of semantic relations. Additionally, LLKBs also benefit the task of sense clustering, as they offer more flexibility and tuning possibilities, and can also be used to produce multi-layered sense-annotated corpora for evaluation and/or training purposes.

Chapter 5 examines further the potential of LLKBs and considers advanced disambiguation methods such as distant supervision and continuous vector space models

of knowledge bases. Distant supervision corresponds to leveraging LLKBs to automatically generate annotated data, which can afterwards be used in semi-supervised machine learning or rule-based approaches. In this setting, the usage of LLKBs is again beneficial, although not yet entirely explored. As for new types of sense representation (knowledge base and sense embeddings), this area of work is still in its infancy and is evolving quickly; here the chapter presents the potentials and reviews recent work.

Chapter 6 considers the usage and benefits of LLKBs in multilingual applications, particularly multilingual semantic relatedness and computer-aided translation. The joint use of information available in multiple languages (such as in BabelNet) leads to a performance boost when calculating the similarity of words across languages, whereas the richness of LLKB information (large coverage, complementary and structured information) provides additional knowledge in translation applications.

Chapter 7 describes how LLKBs can be explored, curated, and used via graphical interfaces and application programming interfaces. These have evolved with the increasing complexity of (L)LKBs and are of primary importance for efficient search and usage of these resources.

Overall, this book provides a detailed introduction to linked lexical knowledge bases and a timely overview of research on this topic. This area has indeed gained in importance during the last decade and continues to evolve rapidly. In this regard, this book is a helpful entry point to the topic. The introduction of different aspects is very pedagogical and progresses in increasing order of complexity. One can only regret the few concrete examples, especially in Chapters 4 and 5. Along with the presentation of the different challenges and existing solutions, existing work is reviewed each time, which gives the reader the possibility of exploring further a specific point. Although some recent points might soon become outdated (especially regarding sense representations, as acknowledged by the authors), this book fully fulfills its mission of introducing the fundamentals of linked lexical knowledge bases.

Maud Ehrmann is a research fellow at the EPFL Digital Humanities Laboratory (DHLAB) in Lausanne, Switzerland. Her research interests lie in historical document and multilingual natural language processing, with special focus on information extraction, textual content analysis, and knowledge representation. Ehrmann's e-mail address is `maud.ehrmann@epfl.ch`.