Preface

In the last years, researchers from a variety of computer science fields including computer vision, language processing and distributed computing have begun to investigate how collaborative approaches to the construction of information resources can improve the state-of-the-art. Collaboratively constructed language resources (CCLRs) have been recognized as a topic of its own in the field of Natural Language Processing (NLP) and Computational Linguistics (CL). In this area, the application of collective intelligence has yielded CCLRs such as Wikipedia, Wiktionary, and other language resources constructed through crowdsourcing approaches, such as Games with a Purpose and Mechanical Turk.

The emergence of CCLRs generated new challenges to the research field. Collaborative construction approaches yield new, previously unknown levels of coverage, while also bringing along new research issues related to the quality and the consistency of representations across domains and languages. Rather than a small group of experts, the data prepared by volunteers for knowledge construction comes from multiple sources, experts or non-experts with all gradations in-between in a crowdsourcing manner. The resulting data can be employed to address questions that were not previously feasible due to the lack of the respective large-scale resources for many languages, such as lexical-semantic knowledge bases or linguistically annotated corpora, including differences between languages and domains, or certain seldom occurring phenomena.

The research on CCLRs has focused on studying the nature of resources, extracting valuable knowledge from them, and developing algorithms to apply the extracted knowledge in various NLP tasks. Because the CCLRs themselves present interesting characteristics that distinguish them from conventional language resources, it is important to study and understand their nature. The knowledge extracted from CCLRs can substitute for or supplement customarily utilized resources such as WordNet or linguistically annotated corpora in different NLP tasks. Other important research directions include interconnecting and managing CCLRs and utilizing NLP techniques to enhance the collaboration processes while constructing the resources.

CCLRs contribute to NLP and CL research in many different ways, as demonstrated by the diversity and significance of the topics and resources addressed in the chapters of this volume. They promote the improvement of the respective methodologies, software, and resources to achieve deeper understanding of the language, at the larger scale and more in-depth. As the topic of CCLRs matures as a research area, it has been consolidated in a series of workshops in the major CL and artificial intelligence conferences,[3] and a special of issue of the Language Resources and Evaluation journal [1]. Besides, the community produced a number of widely used tools and resources. Examples of them include word sense alignments between WordNet, Wikipedia, and Wiktionary [2–4],[4] folksonomy and named entity ontologies [5, 6], multiword terms [7],[5] ontological resources [8, 9],[6] annotated corpora [10],[7] and Wikipedia and Wiktionary APIs.[8]

Purpose of This Book

The present volume provides an overview of the research involving CCLRs and their applications in NLP. It draws upon the current great interest in collective intelligence for information processing in general. Several meetings have taken place at the leading conferences in the field, and the corresponding conference tracks, e.g. "NLP for Web, Wikipedia, Social Media" have been established. The editors of this volume, thus, recognized the need to summarize the achieved results in a contributed book to advance and focus the further research effort. In this regard, the subject of the book "The People's Web Meets NLP: Collaboratively Constructed Language Resources" is very timely. There is no monograph, textbook or a contributed book on this topic to comprehensively cover the state-of-the-art on CCLRs in a single volume yet. Thus, we very much hope that such a book will become a major point of reference for researchers, students and practitioners in this field.

Book Organization

The chapters in the present volume cover the three main aspects of CCLRs, namely construction approaches to CCLRs, mining knowledge from and using CCLRs in NLP, and interconnecting and managing CCLRs.

---

[3] People's Web Meets NLP workshop series at ACL-IJCNLP 2009, COLING 2010, and ACL 2012

[4] http://www.ukp.tu-darmstadt.de/data/sense-alignment/, http://lcl.uniroma1.it/babelnet/

[5] http://www.ukp.tu-darmstadt.de/data/multiwords/

[6] http://www.ukp.tu-darmstadt.de/data/lexical-resources, http://www.h-its.org/english/research/nlp/download/wikinet.php

[7] http://anawiki.essex.ac.uk/

[8] JWPL (http://www.ukp.tu-darmstadt.de/software/jwpl/), wikixmlj (http://code.google.com/p/wikixmlj/), JWKTL (http://www.ukp.tu-darmstadt.de/software/jwktl/)

Part 1: Approaches to Collaboratively Constructed Language Resources

Collaboratively constructed resources have different forms and are created by means of different approaches, such as collaborative writing tools, human computation platforms, games with a purpose, or collecting user feedback on the Web.

Some of them are constructed by applying Social Web tools, such as wikis, to existing forms of knowledge production. For example, Wikipedia was created through the use of wikis to construct an electronic encyclopedia. In a similar way, Wiktionary was created through the use of wikis to construct a user-generated dictionary. Major research questions in this area of research are: how to utilize a Social Web tool to come up with a useful resource, motivating users to contribute, how to extract the knowledge, quality issues, varied coverage, or incompleteness of the resulting resources.

Further CCLRs result from the purposeful use of human computation platforms on the Web, such as Amazon Mechanical Turk, to perform expert-like or highly subjective tasks by a large number of non-expert volunteers paid for their work. Thereby, a complex task is typically modeled as a set of simpler tasks solved by means of a web-based interface. In other settings, platforms for collaborative annotation by non-paid peers may be used to construct language resources collaboratively. Major research questions in this context are, for example, how to model a complex task in such a way that it is feasible to be solved by non-experts, how to prevent spam, or monetary, quality and labor management issues.

The third approach to the construction of CCLRs by means of crowdsourcing is modeling the data management tasks, such as data collection or data validation as a game. The players of such a game contribute their knowledge collectively either for fun, or for learning purposes. These works address research questions such as how to convert the task into a game, how to motivate players for continuous participation, and how to manage the quality of the resulting data.

Part 2: Mining Knowledge from and Using Collaboratively Constructed Language Resources

Much effort have been put into utilizing CCLRs in various NLP tasks and demonstrating their effectiveness. The present volume includes a number of examples for research works in this area, specifically, construction of semantic networks, word sense disambiguation, computational analysis of writing, or sentiment analysis.

The first approach to mining knowledge from CCLRs is to construct or improve semantic networks. There exist manually constructed semantic resources such as WordNet and FrameNet. Resources constructed through collective intelligence such as Wikipedia, Wiktionary, and Open Mind Common Sense[9] can provide rich and real-world knowledge at large scale that may be missing in manually constructed

[9] http://openmind.media.mit.edu

resources. In addition, combining resources that are complementary in coverage and granularity can yield a higher quality resource.

The second approach to utilizing CCLRs is mining the vast amount of user-generated content in the Web to create specific corpora which can be used as resources in computational intelligence tasks. Much of this data implicitly carries semantic annotations by users, as the corpora typically evolve around a certain domain of discourse and therefore represent its inherent knowledge structure. NLP applications exemplified in this book include the computational analysis of writing using Wikipedia revision history, organizing and analyzing consumer reviews, and word sense disambiguation utilizing Wikipedia articles as concepts.

The applications of CCLRs in NLP are certainly not limited to the example topics explained in this book; one can find a large number of research works with similar goals and approaches in the literature.

Part 3: Interconnecting and Managing Collaboratively Constructed Language Resources

Readily available technology and resources such as Amazon Mechanical Turk and Wikipedia have lowered the barriers to collaborative resource construction and its enhancements. They also have led to a large number of sporadic efforts creating resources in different domains and with different coverage and purposes. This often results in resources that are disparate, poorly documented and supported, with unknown reliability. That is why the resources run the risk of not extensively being used by the community and can therefore disappear very quickly.

The research question is then how to create linguistic resources, expert-built and collaboratively constructed alike, more sustainable, such that the resources are more usable, accessible, and also easily maintained, managed, and improved.

In this part of the book, a number of ongoing community efforts to link and maintain multiple linguistic resources are presented. Considered resources range from lexical resources to annotated corpora. The chapters of the volume also introduce special interest groups, frameworks, and ISO standards for linking and maintaining such resources.

Target Audience

The book is intended for advanced undergraduate and graduate students, as well as professionals and scholars interested in various aspects of research on CCLRs.

Acknowledgements We thank all program committee members who generously invested their expertise and time for providing constructive reviews. This book would not have been possible without their support, especially considering the tight schedule and multiple review rounds. We also thank Nicoletta Calzolari for her insightful and inspiring foreword.

Darmstadt, Germany Iryna Gurevych and Jungi Kim

Program Committee

Iñaki Alegria
Chris Biemann
Erik Cambria
Jon Chamberlain
Christian Chiarcos
Johannes Daxenberger
Ernesto William De Luca
Gianluca Demartini
Judith Eckle-Kohler
Nicolai Erbs
Oliver Ferschke
Bilel Gargouri
Catherine Havasi
Sebastian Hellmann
Johannes Hoffart
Inas Mahfouz
Michael Matuschek
Gerard de Melo
Christian M. Meyer
Rada Mihalcea
Tristan Miller
Günter Neumann
Alessandro Oltramari
Simone Paolo Ponzetto
Michal Ptaszynski
Martin Puttkammer
Ruwan Wasala
Magdalena Wolska
Jianxing Yu
Torsten Zesch

References

1. Gurevych I, Zesch T (2012) Special issue on collaboratively constructed language resources. Language resources and evaluation. Springer, Netherlands
2. Niemann E, Gurevych I (2011) The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and WordNet. In: Proceedings of the international conference on computational semantics (IWCS), Oxford, UK, pp 205–214
3. Meyer CM, Gurevych I (2011) What psycholinguists know about chemistry: aligning Wiktionary and WordNet for increased domain coverage. In: Proceedings of the 5th international joint conference on natural language processing (IJCNLP), Chiang Mai, Thailand, Nov, pp 883–892
4. Navigli R, Ponzetto SP (2010) BabelNet: building a very large multilingual semantic network. In: Proceedings of the 48th annual meeting of the association for computational linguistics (ACL), Uppsala, Sweden, July, pp 216–225
5. Tomuro N, Shepitsen A (2009) Construction of disambiguated folksonomy ontologies using Wikipedia. In: Proceedings of the 2009 workshop on the people's web meets NLP: collaboratively constructed semantic resources, Suntec, Singapore, August, pp 42–50
6. Shibaki Y, Nagata M, Yamamoto K (2010) Constructing large-scale person ontology from Wikipedia. In: Proceedings of the 2nd workshop on the people's web meets NLP: collaboratively constructed semantic resources, Beijing, China, August, pp 1–9
7. Hartmann S, Szarvas G, Gurevych I (2011) Mining multiword terms from Wikipedia. In: Pazienza MT, Stellato A (eds) Semi-automatic ontology development: processes and resources. IGI Global, Hershey, pp 226–258
8. Meyer CM, Gurevych I (2011) OntoWiktionary – constructing an ontology from the collaborative online dictionary Wiktionary. In: Pazienza MT, Stellato A (eds) Semi-automatic ontology development: processes and resources. IGI Global, Hershey, pp 131–161

9. Nastase V, Strube M, Börschinger B, Zirn C, Elghafari A (2010) WikiNet: a very large scale multi-lingual concept network. In: Proceedings of the 7th international conference on language resources and evaluation (LREC), Valletta, Malta, May, pp 19–21
10. Chamberlain J, Kruschwitz U, Poesio M (2009) Constructing an anaphorically annotated corpus with non-experts: assessing the quality of collaborative annotations. In: Proceedings of the 2009 workshop on the people's web meets NLP: collaboratively constructed semantic resources, Suntec, Singapore, August, pp 57–62

http://www.springer.com/978-3-642-35084-9

The People's Web Meets NLP
Collaboratively Constructed Language Resources
Gurevych, I.; Kim, J. (Eds.)
2013, XXIV, 378 p., Hardcover
ISBN: 978-3-642-35084-9