Book Reviews

Biomedical Natural Language Processing

Kevin Bretonnel Cohen and Dina Demner-Fushman
(University of Colorado School of Medicine, and National Library of Medicine)

John Benjamins Publishing (Book series on Natural Language Processing,
edited by Ruslan Mitkov, volume 11), 2014, 160 pp; hardbound,
ISBN 978-90-272-4997-5

Reviewed by
Jin-Dong Kim
Database Center for Life Science

The book begins with a declaration that "the intended audience of the book is natural language processing specialists who want to move into the biomedical domain." It is indeed a great introductory textbook to the field of biomedical natural language processing (NLP), particularly for NLP specialists. Browsing the contents, many NLP specialists will find the titles of chapters familiar: "Named Entity Recognition," "Relation Extraction," and so on. Those familiar topics are reformulated in the context of biomedical informatics, with rich biomedical examples and a good amount of explanation.

One of the most important characteristics of the book is that it is very easy to read. Because biomedical NLP is an interdisciplinary area, the quality of this kind of authoring counts on the depth of authors' insight into the relevant subject areas, which include NLP, bioinformatics, medical science, and linguistics. The level of clarity and conciseness throughout the book demonstrates that the authors have a clear and deep understanding of these areas. Complex technical details are successfully distilled into abstract ideas which are at a level that can be understood without much prior knowledge. It is an obviously important feature of an introductory textbook. Because of the ease of reading, I think the book is a good introductory resource even for non-NLP specialists (e.g., bioinformaticians who want to make use of NLP resources).

Another notable characteristic of the book is that bio- and medical NLP are equally importantly described. Although biomedical NLP is often referred to as one subject area, the community has been quite divided into bio- and medical NLP, and it is in fact a rare chance to learn about both at the same time. The importance of understanding both is very well described in Chapter 2.3, "Clinical Text Mining," in the context of translational research.

In following sections, I add more detailed reviews on individual chapters. For convenience of review, I take the liberty of dividing the chapters into three parts.

1. Part I. Introductory Materials (Chapters 1 and 2)

Simply speaking, for NLP specialists, entering into biomedical NLP means a change of the type of texts that are going to be dealt with. In this sense, Chapter 1, "Introduction to Natural Language Processing," is an ideal starting point, presenting the characteristics

doi:10.1162/COLI_r_00281

of various types of biomedical texts and the language in them. Chapter 2, "Historical Background," presents a concise history of the field. Seamlessly interweaving relevant parts of the history of biomedical informatics and biomedical NLP, it is a masterpiece of introduction to the field—if I were asked to recommend only one chapter of the book to read standing between bookshelves of a library, I would pick this chapter.

In the book, explanations about biomedical information often go one step further than I would expect in similar writing. For example, in Chapter 2.4, "Types of Users of Biomedical NLP Systems," model organisms are explained as major targets of database development. The explanation includes the reason why some organisms (e.g., *C. elegans* and zebrafish [*D. rerio*], have been chosen as model organisms. It may be a small addition, but I believe these additional explanations on bioinformatics will make a significant difference to the reader's understanding of biomedical NLP.

2. Part II. Fundamental Tasks (Chapters 3–6)

As in NLP for other domains, named entity recognition, relation extraction, information retrieval, and concept normalization are fundamental tasks of biomedical NLP. Chapters 3–6 describe these four fundamental tasks, in which beginners of biomedical NLP will be immediately interested. Each chapter typically begins with a task definition in general NLP, which is followed by characteristic features specific to the biomedical context. As already mentioned in the previous section, rich explanation about biomedical terms is a strength of the book. For example, Chapter 3.3, "Why Gene Names Are the Way They Are," will deepen the readers' understanding of gene names one step further.

For each task, several software systems are described to exemplify different ideas of how problems of the task can be approached. Instead of briefly mentioning many systems, the authors have chosen to explain a number of representative systems in in-depth detail, which helps the readers to clearly understand the ideas of the approaches.

The authors take a scientific approach to developing discussions. In many places, while describing the tasks, naturally occurring questions are captured, for example, *how useful are gazetteers for named entity recognition?*, *how useful are co-occurrences for information extraction?*, and so on. Then, discussions on how these questions have been addressed follow. This approach is effective in that it not only keeps the readers interested, but also enables the readers to develop a systematic view of the field, to be able to answer questions such as *what are the important problems?*, *what problems are solved, and how?*, and *what problems are yet to be solved?*

3. Part III. Advanced Topics (Chapters 7–11)

Chapter 7, "Ontologies and Computational Lexical Semantics," describes UMLS and GO as important lexical and ontological resources in biomedical NLP, and also describes contribution of NLP techniques to the development of ontologies. One characteristic of the book is that important resources are described repeatedly in different contexts. For example, UMLS, a compendium of many controlled vocabularies in the biomedical sciences, is referred to in five chapters (Chapters 3–7). This underscores the importance of the resource. The extensive explanation of UMLS, in various application cases, is one important contribution of the book. For biomedical ontologies, OBO (Smith et al. 2007) and BioPortal (Whetzel et al. 2011) are also important community resources. Readers are thus recommended to further read about them.

Chapter 8, "Summarization," and Chapter 9, "Question-Answering," describe the tasks as the titles imply. Even for the readers who are not interested in these tasks themselves, the two chapters offer a good chance to learn about various user requirements and evaluation of the domain. For those who are interested in the question answering task, BioASQ (Tsatsaronis et al. 2015) is a recent important community effort, and further reading about it is recommended, too.

Chapter 10, "Software Engineering," is focused on software testing. One may think that the topic is a bit out of the focus for the book. However, considering that software testing is an important but often neglected issue, I think it is an appropriate choice for addressing the topic in an introductory textbook; otherwise, it may continue to be neglected. The chapter begins with an episode that showcases the importance of software testing in the field, which is very effective in motivating the readers to engage in software testing. From the perspective of software engineering, considering that the number of open source projects and interconnection of open Web services are increasing, software testing will become an increasingly important issue.

The book ends with Chapter 11, "Corpus Construction and Annotation," where the authors attempt to identify important features of successful corpora and annotation.

4. Conclusion

I highly recommend the book not only to everyone new to biomedical NLP, but also to those who are already working in it. The level of understanding and vision of the authors is absolutely outstanding, and the book offers a good chance to acquire it. Through the book, I believe the readers will be able to get a sense of the big picture for biomedical NLP, and to understand motivations and requirements from the field of bioinformatics and medical science, as well as common problems and solutions that have to be addressed and developed.

References

Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. 2007. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255.

Tsatsaronis, George, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Whetzel, P. L., N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. 2011. BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39:W541–5.

Jin-Dong Kim is a project associate professor of Database Center for Life Science (DBCLS) in Japan. He is one of the leading authors of the GENIA resources and also a regular organizer of the BioNLP Shared Task series. His research interests include knowledge acquisition and representation with a focus on corpus engineering and annotation. Kim's e-mail address is `jdkim@dbcls.rois.ac.jp`.