Book Reviews

Text Mining for Biology and Biomedicine

Sophia Ananiadou and John McNaught (editors)
(University of Manchester and UK National Centre for Text Mining)

Boston and London: Artech House, 2006, xi+286 pp; hardbound, ISBN 1-58053-984-X, £53.00

Reviewed by
Nikiforos Karamanis
University of Cambridge

Text mining is defined by Hearst (1999) as the automatic discovery of new, previously unknown information from unstructured textual data. This is often seen as comprising three major tasks: information retrieval (gathering relevant documents), information extraction (extracting information of interest from these documents), and data mining (discovering new associations among the extracted pieces of information).

Most researchers in the natural language processing (NLP) community are familiar with work on information extraction and its subtasks such as noun phrase chunking, named entity recognition, and anaphora resolution, typically applied to newswire articles. The explosive growth of biomedical literature has prompted increasing interest in applying such techniques to biomedical text in order to address the information overload faced by domain experts. This is reflected by the proliferation of articles reviewing this work (Reviews 2006), which typically appear in bioinformatics journals and target experts in biosciences as their primary audience.

Text Mining for Biology and Biomedicine provides an overview of the fundamental approaches to biomedical NLP in more depth than is typically offered in a review article. The book consists of an introductory chapter written by the editors and nine chapters that each discuss a different sub-area of biomedical NLP. Each chapter is authored by researcher(s) with significant contributions to the overviewed sub-area.

In the introductory chapter, Sophia Ananiadou and John McNaught adhere to the definition of text mining by Hearst (p. 1), although their interpretation lays more emphasis on the unstructured nature of the input textual data than on the potential novelty of the output information. Those “interested in organizing, searching, discovering, or communicating biological knowledge” (p. 2) are the targeted readers of the book. The book aims to provide them with an “extensive summarization and discussion of the research literature in text mining and reported systems, geared towards informing and educating rather than oriented towards other text mining experts” (p. 3). Information retrieval is placed outside the scope of the book, which focuses on performing familiar tasks such as named entity recognition (NER) and information extraction (IE) on biomedical text, but also extensively discusses problems that are less studied in general NLP but are of particular importance in this area, such as the exploitation of domain-specific knowledge sources, the construction of terminologies, how to deal with abbreviations, and so on. An outline of the main aims and challenges in biomedical NLP is followed by an overview of how these issues are discussed in each chapter.

Chapter 2, “Levels of natural language processing for text mining” by Udo Hahn and Joachim Wermter, first explains how each level of linguistic analysis (i.e., morphology, syntax, and semantics) is associated with distinct NLP components. Then a reference architecture for text mining that combines these components with each other and with domain resources is presented and compared with the organization of two extant systems, GeneWays (Rzhetsky et al. 2004) and PASTA (Gaizauskas et al. 2003). The comparison leads to the conclusion that the reference architecture represents a somewhat "idealized view of system building ... which has to be supplemented by many heuristic solutions" (p. 37), which might explain why this architecture is not referred to extensively in subsequent chapters. The archetypal text-mining system should ideally strive to produce "some form of proposition" (p. 33) as its output, which will be subject to subsequent processing, for example, to discover new knowledge. However, the authors acknowledge that they are not aware of any system with such an advanced reasoning functionality (p. 34). I did not spot such a system being reviewed anywhere else in the book so I may rather safely claim that the book is mainly about conducting NLP in the biomedical domain rather than discussing the text mining process as a whole.

This chapter is meant to serve as "an introduction to the general techniques of NLP ... necessary to fully appreciate discussions in following chapters" (p. 7). The authors cite and discuss the seminal literature for each NLP component in their hypothetical architecture quite comprehensively, although they do not include any references to other introductory readings to NLP. These could have been helpful, because one often comes across terminology that might be unknown to the non-specialist (e.g., Section 2.3.1 on part-of-speech tagging contains terms such as *seed tagging, second order n-gram Markov models, probabilistic suffix analysis,* and *smoothing by linear interpolation* without explanation). Quite a bit of jargon is used in other chapters as well, so my feeling is that the book will be more accessible to readers with some familiarity with NLP than to the non-initiated.

Chapter 3, "Lexical, terminological, and ontological resources for biological text mining" by Olivier Bodenreider, discusses how the major publicly available knowledge sources may support biomedical NER and IE, with particular emphasis on the three components of the Unified Medical Language System (UMLS), namely, the Specialist Lexicon, the Metathesaurus, and the Semantic Network (Bodenreider 2004). The chapter exemplifies the utility that these resources may provide, although Bodenreider also points out that they might often need to be extended or re-engineered to better serve NER and IE. This is one of the main insights of the book that will be repeated in subsequent chapters.

This chapter provides a clear discussion of the commonalities, differences, and complementarity of the existing resources, which are classified into three types: lexical, terminological, and ontological. Not distinguishing between the three types of resources has been claimed earlier in the book to be likely to "lead to confusion and hamper attempts at exploitation for text mining" (p. 7) although, in Bodenreider's own words, this distinction often ends up being "somewhat arbitrary" (p. 55). Hence he appears to concentrate more on some of the limitations of the resources, such as the restricted coverage of the genomic and the molecular biology subdomain. An overview of suggested solutions to this problem is provided, although these seem to be more applicable to domain-specific resources (e.g., model organism databases) than to the more general UMLS resources that the chapter focuses on. It has also been argued earlier that the lack of an explicit link between a lemma in the Specialist Lexicon and the corresponding concept in the Metathesaurus might also limit their utility for NER and IE (p. 27), but this issue remains unaddressed in Chapter 3.

Chapter 4, "Automatic terminology management in biomedicine" by Sophia Ananiadou and Goran Nenadic, focuses on automatic term recognition (ATR) and automatic term structuring (ATS). ATR identifies lexical units that correspond to domain concepts, and ATS organizes the recognized terms into knowledge structures (terminologies). A brief introduction to terminology construction is followed by a presentation of terminological resources in biomedicine (somewhat overlapping with material in the previous chapter). A detailed review of the main approaches to ATM and ATS constitutes the core of the chapter. Equally interesting is the discussion of the challenges posed to ATR by the pervasive phenomena of term variation and ambiguity. The chapter concludes with an overview of the ATRACT system (Mima, Ananiadou, and Nenadic 2001), a terminology management workbench incorporating modules for ATR and ATS. There is much valuable material in this chapter, although I felt that it would have been more appropriate to discuss the difference between ATR and NER here rather than having to wait until Chapter 6 (or go back to page 8). Some discussion of the differences between ATS and IE (overviewed in Chapter 7) would have been useful as well.

Chapter 5, "Abbreviations in biomedical text" by Jeff Chang and Hinrich Schütze, deals with the problem of linking an abbreviation to its expanded form(s). This is important because of the very frequent use of abbreviations in the biomedical genre and the continuous introduction of many new abbreviations. An introductory discussion of the problems of defining and identifying abbreviations is followed by a detailed review of the methods used to construct and evaluate the Stanford Biomedical Abbreviation Database (Chang, Schütze, and Altman 2002). Different types of abbreviation variations (already introduced in the previous chapter, although no cross-reference is provided) and the methods used for their normalization are also overviewed. The chapter also touches upon the problem of identifying long forms that do not appear in the same document as the abbreviation. Several directions for future work are proposed, the most interesting of which, in my view, are the need for a comprehensive study to compare the coverage and accuracy of different abbreviation databases and the more extended investigation of algorithms that can automatically generate abbreviations from long forms.

Chapter 6, "Named entity recognition" by Jong Park and Jung-jae Kim, concentrates on applying NER to biomedical text. The nature of candidate named entities (NEs) and issues related to their ambiguity, variation, and growth rate (also mentioned in previous chapters) are discussed in detail to exemplify how biomedical NER differs from traditional NER in the newswire domain. The main approaches to biomedical NER are reviewed in depth with particular emphasis on the reported evaluation results (although the authors also point out that these cannot always be used to directly compare the approaches because of important methodological differences between the evaluation studies). Grounding the recognized NEs in an ontology and dealing with NEs other than gene and protein names are identified as the main challenges to address in forthcoming research.

Chapters 4 through 6 clearly complement one another, although identifying overlapping or related sections often requires some effort on behalf of the reader. In particular, I often found it hard to keep track of literature reviewed in different chapters. Each chapter contains its own list of references (enumerated in the order in which they appear in the text, which seems to be the norm in biomedical publications). The number in the list is typically used to point to a reference in the text. For instance, the ABGENE system (Tanabe and Wilbur 2002) is reviewed both in Chapter 4 (p. 77) and in Chapter 6 (p. 134). In Chapter 4 the system is mentioned by name followed by its citation number (i.e., "ABGENE [36]"), whereas in Chapter 6 the authors' names appear together with the citation number (i.e., "Tanabe and Wilbur [26]"). (Note that the term ABGENE is not included in the book's index.) Had the book come with a single reference section and a citation index, the reader's attempt to identify and extract related information would have been facilitated greatly.

Chapter 7, "Information extraction" by John McNaught and William Black, is devoted to rule-based methods for the extraction of simple facts and more complex events. An overview of IE as shaped by the MUC evaluation efforts is followed by a comprehensive critical assessment of the various approaches adopted for IE in the biomedical domain. Sublanguage-driven systems that simultaneously consider syntax and semantics, such as GENIES (Friedman et al. 2001), and systems that take advantage of ontological information (Gaizauskas et al. 2003; Cimiano, Saric, and Reyle 2005) are proclaimed to be the most successful. As in Chapter 2, being able to deliver abstract representations of facts and events that can be subjected to subsequent data mining or integrated in a knowledge base to enable reasoning (instead of simply returning textual strings or their transforms) is regarded as a bonus for a system. Given that these representations are likely to be heavily reliant on the requirements of the mining or reasoning process, I would welcome more discussion on how systems developed to deliver material suitable for different knowledge bases may be compared with each other.

Chapter 7 concludes with a call for further efforts to produce resources that can be used to train and evaluate more advanced IE systems, echoing other similar statements throughout the book (most notably in Chapters 2 and 3). Given that the preparation of such resources is not a trivial task (as discussed in the following chapter), it is somehow surprising that semi-supervised or unsupervised machine learning methods, for example those discussed by McCallum (2005), are not mentioned as an alternative research avenue. The problem of resolving anaphoric references is mentioned in several chapters as another essential NLP task that awaits in-depth investigation in the biomedical domain. Chapter 7 is meant to provide an overview of existing approaches to anaphora resolution in biomedical text (p. 148), but this takes place only in passing. Devoting some more space to this issue would have been worthwhile as well.

Chapter 8, "Corpora and their annotation" by Jin-Dong Kim and Jun'ichi Tsujii, discusses issues related to the collection and annotation of corpora. From their own experience in the development of the GENIA corpus (Kim et al. 2003), the authors provide practical advice on how to compile a representative corpus, prepare annotation schemes and guidelines, perform the actual annotation and, ultimately, assess the reliability of the produced data. There is a section on annotation format that lays emphasis on XML-based schemes but does not mention the B-I-O notation that is used in Chapters 2 and 4. An informative discussion on available annotation tools concludes the chapter, which is written very clearly, although I found some material too low level (particularly the script to retrieve MEDLINE abstracts in Figure 8.1) or even subpar (Section 8.3.3 on the comparison of corpora).

Chapter 9, "Evaluation of text mining in biology" by Lynette Hirschman and Christian Blaschke, begins by explaining how the MUC and TREC evaluation challenges and similar efforts in molecular biology inspired community attempts to build shared assessment resources and agree on evaluation methodologies to appraise the state of the art in biomedical NLP. The authors address the key issues of why, how, and what to evaluate and then detail the design, organization, and main results of four recent evaluation challenges and how these should motivate additional efforts in the years to come.

One of the main points in Chapter 9 is that research on biomedical NLP should be focused on resolving problems of practical relevance to biologists in order for them to become involved in the development effort and continue to participate in challenging evaluations. This issue is of particular interest to me because of my current involvement in a project aiming to integrate an NLP system into an existing curation workflow (Karamanis et al. 2007). In Section 9.3.1, the authors distinguish between the tasks performed by different types of users, namely database curators and research scientists, and then go on to explain how different evaluation tasks were designed with a different type of user in mind. Meeting users' requirements seems to be relevant to other chapters of the book as well. For instance, it is not clear whether delivering abstract representations of facts (as suggested in Chapters 2 and 7) will assist curators more than pointing them to actual textual strings. Developing and evaluating integrated systems that address the users' real-world needs is one of the greatest challenges in biomedical NLP (Cohen and Hersh 2005), but is not substantially covered in the book.

Chapter 10, "Integrating text mining with data mining" by See-Kiong Ng, demonstrates how certain NLP techniques may be incorporated into extant algorithms for analyzing nontext biological data such as genomic sequences and expression profiles. Many of the overviewed techniques have been shown to improve solutions to problems that are of particular importance to research scientists, such as homology search and sequence-based functional classification. This is a very interesting chapter, which comes closer to the Hearstian notion of text mining than previous chapters, although most of the reviewed techniques treat the text as a bag of words, thus deviating significantly from the NLP technology discussed previously.

In conclusion, I believe that each chapter successfully combines a comprehensive summary of the fundamental approaches with the authors' precious insights. The book is recommended to anyone interested in a more detailed overview of biomedical NLP than what is typically presented in a review article although, unavoidably, the most recent of these reviews may include more up-to-date information. Readers with some NLP background will probably find the book more easily accessible than biomedical scientists and might benefit even more from it if they apply some additional effort to synthesize opinions across chapters. This review partially reflects my attempt to do that, indicating a few ways in which I think that the book could be further enhanced. In any case, the book will almost certainly fertilize ongoing research in the rapidly expanding area of biomedical NLP, so I felt that studying it was time well spent.

References

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Chang, J., H. Schütze, and R. Altman. 2002. Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Information Association*, 9:612–620.

Cimiano, P., J. Saric, and U. Reyle. 2005. Ontology-driven discourse analysis for information extraction. *Data and Knowledge Engineering*, 55(1):59–83.

Cohen, A. and W. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71.

Friedman, C., P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. 2001. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17:S74–S82.

Gaizauskas, R., G. Demetriou, P. Artymiuk, and P. Willet. 2003. Protein structure and information extraction from biological texts: The PASTA system. *Bioinformatics*, 19(1):135–143.

Hearst, M. 1999. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10, College Park, MD.

Karamanis, N., I. Lewin, R. Seal, R. Drysdale, and E. Briscoe. 2007. Integrating natural language processing with FlyBase curation. In *Proceedings of PSB 2007*.

Kim, J. D., T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA: A semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:i180–i182.

McCallum, A. 2005. Information extraction: Distilling structured data from unstructured text. *ACM Queue*, 3(9):48–57.

Mima, H., S. Ananiadou, and G. Nenadic. 2001. The ATRACT workbench: Automatic term recognition and clustering for terms. In *Proceedings of the 4th International Conference on Text, Speech and Dialogue*, pages 126–133.

Reviews. 2006. *Reviews on Text Mining in Biomedicine*. Biomedical Literature and Text Mining Publications, accessed: 10/10/06. http://blimp.cs.queensu.ca/cateR_1.html.

Rzhetsky, A., I. Iossifov, T. Koike, M. Krauthammer, et al. 2004. GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53.

Tanabe, L. and J. Wilbur. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18:1124–1132.

Nikiforos Karamanis holds a PhD in Informatics from the Institute for Communicating and Collaborative Systems at the University of Edinburgh. Currently he is working as a Research Associate in the BBSRC-funded Flyslip project, a joint effort at the University of Cambridge between the Natural Language and Information Processing Group in the Computer Laboratory and the FlyBase curation team in the Department of Genetics, aiming to integrate NLP tools with manual methods for literature curation (visit www.cl.cam.ac.uk/users/av308/Project_Index/ for more information). His e-mail address is Nikiforos.Karamanis@cl.cam.ac.uk.