Information Extraction: Algorithms and Prospects in a Retrieval Context

Marie-Francine Moens
(Katholieke Universiteit Leuven)

Springer (Information retrieval series, edited by W. Bruce Croft), 2006, xiii+246 pp;
ISBN 978-1-4020-4987-3, $119.00

Reviewed by
Diana Maynard
University of Sheffield

Published as part of the Information Retrieval Series, this book aims to present both a historical overview of information extraction (IE) and a description of current approaches and applications. Essentially, it introduces the topic of information extraction to an information retrieval (IR) audience, aiming as much at students as established researchers, and focusing primarily on the algorithms used. Although it claims to give equal importance to early technologies developed in the field and to the "most advanced and recent technologies," for the latter it concentrates mainly on machine-learning approaches rather than knowledge-engineering (rule-based) techniques.

The book is well-structured and progresses from an introductory chapter that gives a brief explanation of information extraction and how it fits into the IR paradigm, through a historical overview of the field in Chapter 2, and on to the meat of the book, which describes different approaches to IE in a set of four chapters leading from symbolic techniques and pattern recognition through to supervised and then unsupervised classification techniques. Chapter 7 discusses information retrieval models and makes some suggestions as to why and how information extraction should be incorporated into these models. Chapter 8 then discusses evaluation of IE technologies, focusing on the most commonly used measures from the Message Understanding Conference (MUC) and Automatic Content Extraction (ACE) competitions, and suggesting other ways in which evaluation might be measured. Chapter 9 describes some case studies in which information extraction is commonly used. Finally, the author takes a look at the future of information extraction within an information retrieval context, discussing some of the findings and challenges for current research.

The aim of the book is quite ambitious in attempting to cater to the rather different needs of both students and established researchers. On the one hand, the chapters on the history of information extraction would be interesting for students; on the other hand, this could be of lesser interest to researchers concerned about the techniques and applications. The first half of the book is quite readable; the second half, however, targets those with quite some knowledge already of statistical processing, language modeling, and so on. Although the book is designed more for reading right through than dipping into individual chapters, the middle sections are certainly not easy reading, except perhaps for those already familiar with the topic.

It is slightly disappointing that the book is angled almost exclusively towards machine-learning approaches to IE, and consequently omits any mention of some of the leading tools in the field such as GATE (Cunningham et al. 2002) and KIM (Popov et al. 2004), which are based on knowledge-engineering approaches. Although the book does not concern itself with examining individual IE tools, one might expect some reference to such tools, at least in the historical chapters if nowhere else. It is also unfortunate that the quality of the English is quite poor: The dozens of spelling and grammatical mistakes impair the quality and, at times, readability of the work.

The first two chapters of the book, introducing the topic and detailing the history of IE, are most interesting and present an overview of the development of the field that is hard to find in other work. One might have expected to find more about the MUC competitions in this section, which shaped the way for much current research in IE by pushing the technology towards real applications and providing a mechanism by which future tools could be evaluated and compared. There is some discussion of the MUC evaluation methodology later in the book, however. The chapter on symbolic techniques is also written very much from a historical perspective, but does not seem to contribute much to the book as a whole, and is probably of most interest to linguistics students. The chapter on pattern recognition, which one might expect to deal with approaches for rule-based IE, focuses more on investigating the different information units and features that are used for machine-learning techniques, described in the following chapters. Whereas the classification of which features are useful for which tasks is most useful, the omission of more discussion about hand-crafted approaches is a little disappointing. The section on active learning also has some omissions. There is no reference to some of the earliest work such as that of Thompson, Califf, and Mooney (1999), and the author's claim that experiments with such techniques are "recent and limited" is perhaps a little too dismissive.

The chapters on supervised and unsupervised classification are geared towards describing different machine-learning methods with particular respect to IE tasks such as named entity recognition, and give a very detailed overview of the approaches possible. A useful addition would have been a summary of the pros and cons of each method: The reader has to delve deep into each section in order to find this information. The material is well described but still quite difficult to follow for someone without a statistics background—it's definitely not a "Machine Learning for Dummies"—and again, more examples would have been useful.

Chapter 7 discusses how the results of IE could be applied to retrieval models. It first describes the IR models used, which is interesting for someone unfamiliar with the field. However, the discussion of the integration of IE does not seem convincing and there is once again a lack of real examples which would strengthen the arguments. Chapter 8 is devoted to evaluation techniques for IE. Although the explanation is thorough, it is quite complicated and would have benefited from a discussion of the difference between the approaches and the appropriateness of each for different tasks. The complicated ACE value, as used in the Automatic Content Extraction series (a successor of MUC) is explained in great detail, but even so it is hard for the reader to grasp the intricacies of the model and a real understanding of the benefits of each part of it. Most of the discussion is limited to MUC and ACE measures; however it would have been even more interesting if the author had investigated other possible criteria, which are only briefly mentioned at the end, and with no indication of how to evaluate them. For example, the difficulty of evaluating information extraction with respect to an ontology is not discussed, yet this is a very timely and important issue with the advent of systems that categorize entities into many related classes, rather than the handful of classes traditionally used.

Chapter 9 aims to illustrate the different IE approaches and tasks with a set of applications, looking at six domains in which IE is commonly used: news, biomedicine, criminal intelligence, and business, legal, and informal texts. The focus is more on the needs and problems in these areas than on specific applications that have been developed. Although the chapter gives quite a good overview of these domains, it would have been useful to see some more concrete examples explaining exactly how the use of IE technology would be of use, as is done for the intelligence domain. For example, the section on news texts just describes the MUC and ACE competitions but might leave the uninitiated reader still guessing as to what the technology might practically be used for in this field. Some interesting problems are highlighted but unfortunately there are no real solutions to these as yet: for example, problems in the scalability of IE systems are still a major impediment to the more-widespread uptake of such tools in industry.

The final chapter sums up the most important findings from the previous chapters and makes some suggestions for the future of Information Extraction in a retrieval context. Although it does not offer any earth shattering new ideas here, it does make some very valid points about the importance of relation finding and advanced retrieval techniques.

In summary, the book offers a useful overview of the field of IE for those unfamiliar with the topic, but it is probably not for those seeking novel ideas or an application-oriented approach. It could be an appropriate textbook for those in the IR field looking for an introduction to IE, but probably not as useful for those in the IE field looking to expand their horizons into IR. Overall, I would suggest the book as one for the library rather than the personal bookshelf.

References

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 168–175, Philadelphia, PA.

Popov, Borislav, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. 2004. KIM—A semantic platform for information extraction and retrieval. *Natural Language Engineering*, 10(3/4):375–392.

Thompson, Cynthia A., Mary Elaine Califf, and Raymond J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the 16th International Machine Learning Conference*, pages 406–414, Bled, Slovenia.

Diana Maynard is a Research Associate at the University of Sheffield. She has interests in information extraction, tools for language engineering, terminology, evaluation, and accessibility of technology. Her address is Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello St., Sheffield, S1 4DP, UK; e-mail: diana@dcs.shef.ac.uk.