Mach Translat
DOI 10.1007/s10590-011-9086-9

BOOK REVIEW

Colin de la Higuera: Grammatical inference: learning automata and grammars

Cambridge University Press, 2010, iv + 417 pages

Walter Daelemans

Received: 4 November 2010 / Accepted: 25 January 2011

Grammatical inference (GI) is concerned with the study of algorithms for learning automata and grammars from strings. Colin de la Higuera sees a subtle, informal, and almost philosophical, distinction between grammar induction (finding a grammar that explains as much as possible of the data) and grammatical inference (finding the true or only target grammar covering a set of strings). The latter puts more emphasis on the learning process. Although this field has existed for a long time, almost twenty years with its own conference (International Colloquium on Grammatical Inference, ICGI), this is the first monograph on the subject. It has the aim of bringing together the most salient concepts and results, until now dispersed over different publications, in a uniform notation and framework.

In Chap. 1, the author surveys the scientific areas that have contributed to the field of grammatical inference. Computational linguistics, obviously, with its formal work on language learnability, and inductive inference, which studies questions about what classes of functions can be learned, how fast, and how to measure learning success. There is also pattern recognition and machine learning (especially complexity results in computational learning theory) and computational biology (DNA analysis). The latter is an application area of GI that has contributed to the main results in that field. The introduction also establishes the main research questions and methodology of GI. Strings and trees are ubiquitous in language data, biological data, and computer programs, but also in processed image data, music, chemical data, etc. These data and the potential for GI to work on it are described in Chap. 2. Machine Translation is mentioned in this book only in passing, but some of the techniques described in the book have also been used in this area. For example, Chap. 18 describes in detail the OSTIA system for learning transducers that may have some potential as a model for

W. Daelemans (✉)
CLiPS, University of Antwerp, Antwerp, Belgium
e-mail: walter.daelemans@ua.ac.be

Published online: 12 February 2011

Machine Translation. Note that only learning from strings is covered in detail in this book; inferring tree automata and grammars is postponed to a future book.

The next three chapters (Part I, The Tools) introduce the concepts of strings, languages, distributions over strings, and combinatorics, respectively. Chapter 3 gives a compact overview of concepts related to strings, and how to compare them. It also goes into the details of string distances. It is useful to have all the "stringology"-related notation and definitions in a single chapter. Regular and context-free languages are introduced in Chap. 4, and their probabilistic versions in Chap. 5, as well as distances between distributions. In Chap. 6, we learn about combinatorial properties of automata and grammars, and are introduced to concepts such as VC-dimension (from Vapnik–Chervonenkis) and complexity, and to issues of consistency, the search space for learning, and the problem of showing the equivalence of automata and grammars. One advantage in the context of teaching is that the book consistently uses the same terminology and notation, which is a great help for students, as many different notations abound in different contexts. The disadvantage is that readers will sometimes miss the notation they are used to.

Part II introduces the conceptual framework and constraints of (efficient) language learning. In Chap. 7, Gold's infamous concept of identification in the limit with its negative learnability results is introduced, and various possible resource-bounded versions explained. Noisy data makes the task even more difficult. The approach is applied to learning from text in Chap. 8. Readers interested in what these results mean for human language acquisition, may be slightly disappointed. However, good recent work by Clark and Lappin (2010) can complement the discussion in this book. Chapter 9 introduces the concept of active learning in a formal framework. These days any computational linguist is familiar with this approach in an annotation context. In the mass of data that can be annotated, let a machine learner decide which data points are most complex or surprising compared to what the learner already knows, and pick those to be annotated. To see the concept here integrated in a formal framework (from which it actually originated), and not simply as a pragmatic trick provides more comprehensive understanding. The concept of probably approximately correct learning (PAC-learning), due to Angluin, is applied to learning from text in Chap. 10.

Part III defines and describes the properties of various solutions to the negative results on learnability in Part II by adding extra bias (i.e. by assuming various constraints on the general learning situation). Each time the algorithms are described in pseudocode and variations and shortcomings are clearly described. This part makes the book an excellent reference work for quickly finding information about different types of learning algorithms. In Chap. 11, for example, window languages are described that presume bounded memory. This chapter also describes look-ahead languages, pattern languages, and planar languages, and their associated learning algorithms. Informed learners, learning from positive and negative data, are described in Chap. 12, and Gold's algorithm and RPNI (regular positive and negative inference) algorithms are explained. Another way of adding bias in the learning process is the query and oracle based approach in Angluin's LSTAR algorithm (Chap. 13). The next chapters (14–18) describe heuristic procedures like genetic algorithms, the MDL (minimum description length) principle, Tabu search, etc. Various algorithms for learning context-free grammars, probabilistic automata, probability estimation algorithms and learning transducers are reviewed as well. With Chap. 17 about probability estimation, computational linguists will suddenly feel on known territory again with discussions of expectation maximization, Baum-Welch, and the inside-outside algorithm. Again, it is nice to see these techniques from the perspective of GI, and noticing the connections between different subfields of language learning.

The author has a pleasant writing style and succeeds in finding the right examples and matter-of-factly explanations for concepts that are sometimes tough to understand. The introductory example on the prisoner's dilemma, for example, succeeds in immediately drawing the reader into the way of thinking needed to understand and appreciate the methodology of GI.

I especially liked the sections on further reading at the end of each chapter, as they often also point out interesting alternative lines of research, mention anecdotes, and make you think about the open problems. Although the book is not intended as (only) a textbook, most chapters have exercises that are sometimes rather complex. It would be helpful to make the solutions available somewhere.

Way back in 2000, I organized a CoNLL (Conference on Natural Language Learning), the flagship conference of the special interest group on language learning (SIGNLL) of the ACL (Association for Computational Linguistics), together with ICGI. The goal was to explore what we could learn from each other, because after all we were all targeting language learning from data. The field of Computational Linguistics had by then already moved firmly into the statistical paradigm. More than ten years later, I'm still thinking more bridges are necessary, and this book gives some good examples why. It is clear that there are many connections between approaches used in statistical and machine learning based natural language processing and GI, and I referred to a few in this review. But there are many more: ideas about convergence, complexity, generalization, noise, etc.

This is why the concluding Chapter was my only slight disappointment when reading the book. I had expected a broader synthesis of learning that also included current approaches in machine learning and data mining (as applied in computational linguistics) and defined the place of GI in that field. But I realize that is a very tall order. Only showing some of the connections is already a great result.

The scope of the book, the detail of description, the uniformity of notation and treatment, and the enjoyable style make this book an important addition to the library of any computational linguist interested in language learning from data.

Reference

Clark A, Lappin S (2010) Linguistic nativism and the poverty of the stimulus. Wiley Blackwell, Oxford and Malden