Book Reviews

Corpus-Based Methods in Language and Speech Processing

Steve Young and Gerrit Bloothooft (editors)
(Cambridge University and Utrecht University)

Dordrecht: Kluwer Academic
Publishers (Text, Speech and Language
Technology series, edited by Nancy Ide
and Jean Véronis, volume 2), 1997,
xii+234 pp; hardbound, ISBN
0-7923-4463-4, $75.00, £46.00, Dfl 120.00

Reviewed by
Rebecca Bruce
Southern Methodist University

The past three decades have seen a steady growth of interest in corpus-based techniques for speech and natural language processing. Beginning with the success of the early Markov chain–based speech recognition systems (Jelinek 1998) and continuing with the work done in information extraction as part of the Message Understanding Conferences (http://www.muc.saic.com), corpus-based speech and language processing have flourished. Such techniques are widely considered to be the primary means of developing broad-coverage speech and language systems that can be ported to different domains. As such, the range of corpus-based techniques and their applications have expanded greatly in the past decade.

Corpus-Based Methods in Language and Speech Processing edited by Steve Young and Gerrit Bloothooft is a compilation of works presented during a two-week course at the 2nd European Summer School on Language and Speech Communication held in 1994 in Utrecht. The book promises to offer an in-depth introduction to the field of corpus linguistics that is both accessible to newcomers and valuable to practitioners. While the book represents an interesting attempt to meet this ambitious objective, it falls short in at least two significant ways: (1) it fails to adequately survey the full range of the field; and, more importantly, (2) the material is often presented in a manner that is not tailored to the needs of a newcomer. The book does provide a mathematically rigorous presentation of a core set of primarily statistical techniques in both speech and natural language processing and would be a valuable reference for practitioners.

The material is organized into six loosely coupled chapters, each written by a different set of authors. The notation varies slightly between chapters, as does the style of presentation. Chapter 1 by Hermann Ney illustrates the role of statistics in the context of automated speech and language processing. It is a somewhat loosely structured discussion of such topics as Bayes's decision theory and the EM algorithm in the context of applications such as speech recognition and text translation. The mathematical presentations are formal but lack intuitive appeal and may be tedious and difficult for a newcomer to assimilate. There is also a lack of discussion of basic concepts; for example, the concept of stochastic independence is not explicitly discussed. At the same time, the chapter admirably demonstrates the unity in the formulation of some of the basic methodologies applied to natural language processing.

Chapter 2 is by Kate Knill and Steve Young. It provides an extensive survey of speech recognition methodologies based on hidden Markov models. The survey is thorough, but the presentation is not easy to read and has few illustrative examples. Formulae and descriptions occasionally make use of terms and symbols that are not defined for the newcomer.

Egidio Giachin and Scott McGlashan present a spoken-language dialogue system in Chapter 3. Such a system represents the ultimate application of combined automated speech and language technology. The system described was developed as part of the ESPRIT SUNDIAL project (Peckham 1991) and is a combination of hidden Markov model–based speech recognition and rule-based natural language understanding. The system is presented with adequate clarity and detail to allow appreciation of the problems addressed and the methodologies applied to their solution, although the discussion of competing methodology is very limited.

Part-of-speech tagging and partial parsing are discussed by Steven Abney in Chapter 4. The focus of this chapter is primarily on rule-based systems. The survey is thorough and makes a number of interesting points. For example, Abney notes the impact of a Zipfian distribution on evaluation results when discussing taggers; he also provides a good discussion of the longest-match rule when describing partial parsing.

Chapter 5 by Rens Bod and Remko Scha is a carefully written and enjoyable presentation of the formulation of stochastic tree-substitution grammars from a database of labeled trees. Three approaches of increasing complexity are described and evaluated. The third and final approach makes use of the Good–Turing method for estimating the probability of unseen events.

In the final chapter, Hermann Ney, Sven Martin, and Frank Wessel describe methods for estimating the probability of unseen events in n-gram models. Although mathematically complete, the presentation is difficult to follow and may not be accessible to a typical newcomer.

In summary, the book is well worth reading, although it may not deliver all that it promises. It is a technically sound survey of several of the most widely used techniques in (primarily statistical) corpus-based speech and natural language processing. It also contains an extensive bibliography. Unfortunately, its use as a reference is diminished by the lack of an index.

References

Jelinek, Frederick. 1998. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge.

Peckham, Jeremy. 1991. Speech understanding and dialogue over the telephone: an overview of progress in the SUNDIAL project. *Proceedings of the 2nd European Conference on Speech Communication and Technology*, Genova, September 1991, pages 1469–1472.

Rebecca Bruce is an Assistant Professor at Southern Methodist University. Her research focuses on the development of statistical methods for text processing. Bruce's address is: Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX 75275; e-mail: rbruce@seas.smu.edu