The Balancing Act: Combining Symbolic and Statistical Approaches to Language, edited by Judith L. Klavans and Philip Resnik

Reviewed by
Ted Pedersen
Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
tpederse@csc.calpoly.edu

February 18, 1999

Computational approaches to language processing arose during a time of bitter debate in the linguistics community that pitted the generative theory of grammar [Cho57] versus more quantitative and empirically motivated approaches (e.g., [Sha48]). An unfortunate side effect of this feud was a schism between statistical and symbolic approaches to natural language processing that resulted in limited interaction between the two camps until the 1980's.

Fortunately, much of the animosity has passed and we are now in an era of relatively good feeling where it is widely recognized that most real–world problems in language processing can only be cracked with the combined muscle of statistical and symbolic approaches. *The Balancing Act* provides an insightful overview of this important shift in attitudes and practice.

The eight chapters in this book are based on papers presented at a workshop held at the 32nd Annual Meeting of the Association for Computational Linguistics, on July 1, 1994 in Las Cruces, NM. The objective of the workshop was

> ...to provide a forum in which to explore combined symbolic and statistical approaches in computational linguistics.

In their preface, editors Klavans and Resnik outline the combination of circumstances that resulted in increased interest in such hybrid approaches. They point out that it has become apparent to those working on real–world problems that neither statistical nor symbolic approaches are by themselves sufficient; statistical approaches are robust and achieve broad coverage but lack the insightful domain knowledge that symbolic methods can provide. In addition, we now have available cheap and powerful computing hardware, large amounts of online corpora, and high quality online dictionaries and thesari, all of which make empirical statistical approaches easier to investigate than ever before. The editors also suggest that there is a growing realization of more common ground between the statistical and symbolic approaches than has generally been recognized:

> An obvious fact that is often forgotten is that every use of statistics is based upon a symbolic model.

Indeed, every statistical approach must be based on a set of features that represents the linguistic event being modeled. The degree to which that set of features adequately captures the subtleties of the event is the prime factor in determining success or failure.

The chapters included in this book are well–chosen in that they represent a broad range of perspectives. A reader not familiar with computational linguistics or the history of the statistical–symbolic schism is advised to begin with the preface and then read the chapters by Abney and Price.

Chapter 1, Statistical Methods and Linguistics, by Steven Abney, addresses a long–standing criticism of computational linguistics made by the theoretical linguistics community; it is often charged that computational approaches simply result in descriptive analyses that do not account for the ability of humans to create and invent language. Abney counters this with an argument showing that weighted stochastic grammars can offer insights into complex issues involving change and evolution of language.

Chapter 6, Combining Linguistic with Statistical Methods in Automatic Speech Understanding, by Patti Price, reviews progress in speech understanding achieved by combining statistical speech recognition technology with symbolic approaches to natural language understanding. This chapter provides a particularly good overview of terminology and concepts relevant to both statistical and symbolic approaches and will prove useful as the reader moves into some of the more technical material.

The identification of meaningful word combinations in large amounts of text is a problem where statistical approaches sometime employ little or no linguistic knowledge and rely primarily upon frequency counts of words that occur in close proximity to one another. The counts are evaluated using a variety of association measures, with some combinations being identified as statistically significant. However, it is often the case that these significant collocations provide little insight to the task at hand. Two chapters in this book offer approaches that improve upon this methodology by introducing symbolic knowledge.

In Chapter 3, Study and Implementation of Combined Techniques for Automatic Extraction of Terminology, by Beatrice Daille, a case is made that co–occurrences should be defined not by the proximity of words but rather using shallow syntactic information. This is shown to result in more accurate identification of significant collocations for a terminology bank for a technical domain. Daille also notes that after incorporating linguistic knowledge in the process, simple frequency counts achieve results that are just as useful as those of statistical tests of significance.

Chapter 4, Do We Need Linguistics When We Have Statistics? A Comparative Analysis of the Contributions of Linguistic Cues to a Statistical Word Grouping System, by Vasileios Hatzivassiloglou, also discusses the identification of collocations; this time in order to identify conceptually related groups of adjectives. A novel evaluation procedure is described where a human judge performs the same conceptual grouping as the computational approach; however, the human is essentially unaware that they are being used as a point of comparison and therefore produce a truly independent gold–standard.

Readers with an interest in machine learning should take particular note of Chapter 7, Exploring the Nature of Transformation-Based Learning, by Lance A. Ramshaw and Mitchell P. Marcus. They provide a detailed explanation and analysis of transformation–based learning (e.g., [Bri93], [Bri94]), a corpus–based approach that has been applied to part–of–speech tagging, parsing, and prepositional phrase attachment. The focus of the chapter is on the resistance of transformation–based learning to over–training; this is investigated in depth and numerous comparisons are made to decision tree learners.

Chapter 2, Qualitative and Quantitative Models of Speech Translation, by Hiyan Alshawi, argues that the contrast between statistical and symbolic approaches is often exaggerated. Alshawi correctly points out that many symbolic rule–based frameworks are capable of learning from large corpora while the parameters of statistical approaches can have their values set by intuitive rather than empirical means. Alshawi suggests that the real distinction is between systems that manage constraints and those that compute numerical functions. This is illustrated by a speech translation where some logic based constraints are replaced with statistical associations to good effect.

Chapter 8, Recovering from Parser Failures: A Hybrid Statistical and Symbolic Approach, by Carolyn Penstein Rose and Alex H. Waibel, is also related to speech translation. They focus on the difficult problem of parsing in the face of ungrammatical input, a common occurrence when processing speech rather than written text. Their approach creates a partial parse of such utterances that is then reconstructed based on information obtained during a dialogue with the user.

Chapter 5, The Automatic Construction of a Symbolic Parser via Statistical Techniques, by Shyam Kapur and Robin Clark, describes an approach that defines a space of parameters and locates a target language in that space based on negative evidence contained in unprocessed sentences. This includes a very interesting discussion on the nature of the parameters used to define this space of possible languages, as well as the relationship of this approach to child language acquisition.

Books based on a collection of papers from a workshop are sometimes a frustrating affair; too often they are simply the working notes bound up in a shiny new cover. It is a great pleasure to report that *The Balancing Act* is exactly the opposite. Indeed, the editors are to be commended for the obvious care taken in the preparation of this book. The papers included have gone through additional rounds of reviews that clearly resulted in significant improvements over the original workshop presentations. The editors have taken the additional meritorious step of writing introductions to every chapter; the work therein is outlined and placed in a larger context. In addition, their preface is of considerable help in understanding the historical differences that have existed between the statistical and symbolic communities.

The Balancing Act is highly recommended to readers who have an interest in understanding both the history and current state of statistical–symbolic approaches to natural language processing. Since nearly five years have passed since the original workshop, it would be appropriate for students and researchers to augment this book with more recent papers. A reader with a limited background in statistical language processing would be wise to consult supplementary background material. [Cha93] is recommended as an approachable yet thorough treatment of most of the statistical issues that arise in this book.

References

[Bri93] E. Brill. Transformation-based error-driven parsing. In *Proceedings of the 3th International Workshop on Parser Technologies*, Tilburg, The Netherlands, 1993.

[Bri94] E. Brill. Some advances in rule-based part of speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1994.

[Cha93] E. Charniak. *Statistical Language Learning.* The MIT Press, Cambridge, MA, 1993.

[Cho57] N. Chomsky. *Syntactic Structures.* Mouton, The Hague, 1957.

[Sha48] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3–4):379–423,623–656, 1948.