Book review

The Balancing Act, Judith L. Klavans and Philip Resnik

Diana McCarthy
School of Cognitive and Computing Sciences, University of Sussex,Falmer, Brighton, BN1 9QH. Email: `dianam@cogs.susx.ac.uk`

BOOK TITLE: The Balancing Act
AUTHOR(S) OR EDITOR(S): Judith L. Klavans and Philip Resnik
TOWN OF PUBLICATION: London, England
YEAR OF PUBLICATION: 1996
SERIES:
PUBLISHER: The MIT Press
PRICE IN $ AND/OR £:
PRELIM. PAGES + ARABIC PAGES + INDEXES (COMBINED OR SUBJECT AND/OR AUTHOR'S): 10 + 179 + 4
ISBN: 0-262-61122-8

Until the 1990s statistical and symbolic approaches to NLP had been seen as two quite separate paradigms but in recent years the advantages of combining the two have been recognised (Gazdar, 1996). This book comprises a selection of the papers presented at a 1994 workshop which aimed at promoting discussion of such hybrid systems. Throughout the book the advantages of statistical approaches are reiterated — the ability to produce robust systems with wide coverage and graceful degradation that can handle ambiguity in a tractable manner. These characteristics cannot be achieved by a symbolic system alone. The rationale for using symbolic information is given less emphasis: to ensure the statistics are collected over linguistically motivated data instances, (for example to handle the multitude of situations where the relevant context is not local), to help diminish the inevitable effect of sparse data, and to ensure that any output is meaningful.

The two primary goals of the collection are stated in the preface. The first is to demonstrate that the two paradigms are not contradictory. The second is to investigate the "balancing act" required when the two are combined. At the time of the workshop the two approaches may have been viewed as contradictory by many researchers, yet the claim to the contrary does not seem very controversial at the current time. Exploration into the type of symbolic and statistical models which can complement each other is however extremely worthwhile as the NLP community still probes for answers in this direction. All papers in the book contribute to the motivation for hybrid models and all but the introduction pursue the wide variety of issues that spring from the second goal.

Most computational linguists are interested in solutions to engineering problems but there are those whose endevours are concentrated on modeling some aspect of human cognition. The need to handle "real" data is clearly important for both. Whilst the bulk of the book is oriented towards engineering there are two chapters which reflect on human processing. The first of these, the introduction by Abney, motivates the use of statistics in the field of linguistics proper, a discipline that appears to have been untouched by the tide of change in the computational linguistics field.

Abney argues convincingly that statistics provide not simply a way of overcoming the engineer's lack of knowledge but a mechanism to understand how humans cope with vast amounts of ambiguity and how they learn language when faced with erroneous examples. Language acquisition, change and variation are each characterised by a gradual change along a continuum with shifts in frequencies over time (or location). Thus a statistical framework is advocated to explain why a feature (or language) is not simply in one moment and out the next.

A major reason for the lack of statistical notions in theoretical linguistics stems from a preoccupation with Chomsky's "competence", much of what is traditionally thought of as performance is neglected. Abney argues that this distinction gives rise to a rather unnatural division between grammar and processor. Focusing on neat examples and rare phenomena might make the study of language more manageable but does reveal much of what goes on when humans process language. It is not that we shouldn't look at these interesting examples but that avoidance of "performance" data denies us an understanding of issues central to our ability to handle natural language, for example how we cope with the myriad of possible analyses for a given sentence. Abney insists that handling large pieces of language data is not done by some simple multiplication of the processes that deal with smaller fragments. Unless linguists study this data their theory will be deficient. Abney focuses on the the importance of stochastic grammars for coping with ambiguity and graceful degradation in the face of real data. He also advocates the use of distributional induction methods, for lexical acquisition and handling of unknown words, but these techniques are not given much emphasis in his chapter.

The second chapter, by Alshawi, concentrates on a specific natural language processing application, speech translation. He argues for a qualitative—quantitative distinction rather than the symbolic—

statistic one. [1] The bulk of the chapter highlights the power of harnessing lexical sensitivity. Alshawi contrasts two designs for a speech translation system, in both designs this is a transfer system with no interlingual symbols. The parts of the conventional qualitative system which would benefit from the addition of quantitative methods are highlighted. Principally the qualitative system is unable to rank analyses and requires a difficult compromise to ensure a wide enough coverage whilst avoiding ungrammatical cases. He notes that qualitative systems are typically oriented towards capturing generalisations yet more specific information in the form of lexical collocations would dramatically improve performance. It is these lexical dependencies that his quantitative model picks up on, whilst having a simpler representation than the qualitative one. This quantitative model relies on a statistical dependency grammar which combines both structural and collocational information. The semantics allowed by the translation model are rather simple but we are reminded that even the qualitative model had to be restricted to first order logic in order to be practical. The qualitative and quantitative systems have similar global architecture but the lexical sensitivity of the quantitative model helps to narrow the otherwise vast search space and avoiding hard and fast constraints ensures the system is less brittle. The quantitative model does however use "symbolic" knowledge, relying on a recursive linguistic structure.

The acquisition of lexical collocations, such as those required by Alshawi's system, and indeed acquisition of other lexical information is a large and active area of research in computational linguistics. One research question is how far such acquisition can go without any linguistic guidance. The third and fourth chapters relate to this issue. In the system for terminology extraction described in Daille's chapter linguistic preprocessing is applied to ensure the statistics are gathered from data in specific syntactic relationships. Many other systems also use syntactic preprocessing before observing distributional evidence, for example see (Pereira, Tishby and Lee, 1993; **?**; Grefenstette, 1992) and also the chapter by Hatzivassiloglou in this volume. Daille outlines problems with going the other way round, using linguistic filtering after statistical evidence has been gathered. The regular nature of the syntactic constraints for terminology extraction makes specification of the linguistic filter easy, other applications might require more sophisticated machinery and this is another area where combined approaches may prove fruitful.

[1] The contrast recommended is between an algebraic system and one involving numeric computation. Perhaps the terms qualitative—quantitative reflect the distinction somewhat better but I am unconvinced that the extra jargon is really helpful.

Daille's chapter goes on to compare some statistics for this task. Interestingly mutual information, a statistical measure of association, fairs rather poorly when compared to straightforward frequency data. Mutual information has been criticised for over-estimation for rare events (Dunning, 1993) but on this occasion it is criticised for selecting frozen compounds typical of language generally rather then domain specific terms. Such comparisons of statistical techniques are invaluable to researchers looking for the appropriate statistical tool for a given NLP task. Daille's comparison is undoubtedly helpful to the terminology extraction application but care should be taken before transferring findings about metrics from this task to another.

Hatzivassiloglou's chapter also concerns lexical associations. The application is the classification of adjectives by clustering on the basis of distributional evidence provided by the modified nouns. This chapter concentrates less on the features of the system and more on techniques for evaluating the contribution provided by the various sources of linguistic guidance. A good feature of the evaluation is that the level of disagreement between judges is taken into account by a weighting on the agreement between the judges' decisions. The linguistic knowledge evaluated includes morphological processing of the adjective—noun pairs (this ensures morphological variants of the same root are considered together), spell checking and also various levels of linguistic sophistication in the process that extracts the adjective— noun pairs. Four levels of knowledge are used in the extraction process and these range from the baseline of simply taking all nouns within the same sentence as an adjective to parsing the input using a finite-state grammar. All the above sources of linguistic knowledge help provide "positive" evidence that two adjectives should be clustered together by virtue of having similar distributions with respect to the modified nouns. Additionally a noteworthy use of "negative" evidence is used which indicates when two adjectives should NOT be placed together. This linguistic guidance occurs in cases where two adjectives appear together modifying the same noun. These are never placed in the same group on the grounds that the adjectives must be suppling different information.

All sources of linguistic knowledge, except the spell checker, were found to improve performance significantly. This is so even for a 21 million word corpus. Further work proposed on the effects of linguistic knowledge and corpus size should prove interesting. Hatzivassiloglou makes a relevant comment on the cost of adding these "symbolic" components. The verdict is that the cost is not excessive and certainly does not outweigh the benefits. The negative evidence makes an interesting contribution to this work but it is unclear how easily it will be to identify this type of information in other applications.

A major drawback of existing techniques for distributional clustering is the production of incongruous classes (Resnik, 1993). On the other hand man- made resources have other drawbacks, particularly their lack of tailoring to the domain of interest. Hatzivassiloglou notes both the possibility of post-editing automatically produced classifications as well as suggesting that on the basis of results here it may be worth starting from some lexical semantic knowledge, although he doesn't say what form that might take. It would be interesting to see where human intervention is best exploited, with regard to both quality and cost.

The chapter by Kapur and Clark is the only one other than Abney's which considers human cognition. It concerns language acquisition and is relevant to both cognitive and engineering concerns. It involves setting the parameters required for a traditional symbolic parser automatically using statistics. The interaction of parameters and gradual accumulation of knowledge during the setting is of interest both as a model for child language acquisition and because it makes sense for a parser to build on what it already knows, with the advantage that it will be more robust in the face of new or noisy data. In this scheme the more frequent parameters are set first. These parameters must be set with little known about the target language and rarer parameters can then be acquired from the structures so far established.

A variety of interesting parameters are discussed but the main focus is on the V2 (verb-second word order parameter) and differences between free pronouns and pro-clitics. Distributional evidence is used and such evidence is claimed to be crucial for any theory of parameter settings. In this model distributional evidence is used to measure entropy at contrasted positions with respect to the verb (V2 parameter) or pronoun (identification and classification of clitic pronouns). The statistical nature of the learning allows trigger detection to occur gradually, and the system can tell when it has seen enough data to reliably set a parameter.

Price's chapter concentrates on the application of speech understanding and in particular on differences between speech and natural language processing. It reiterates background information on the bias of the speech and language communities towards statistical and symbolic approaches. Price argues that the integration of symbolic and statistical approaches is a matter of culture difference between these two communities rather than a technical difficulty of combining these two paradigms. The bias of psychologists and linguists to a symbolic framework and engineers to a statistical one is also mentioned. Price's overview mentions components and technologies of both speech recognition and natural language understanding. She summarises current methods in both and gives an indication of how integration of these two disciplines would give the advantage of added constraints between the different levels. She ends her chapter with a comment on the challenges that she feels need to be addressed for integration. For the speech community she suggests addressing the notion of a prototype and distance from a prototype would increase robustness. She feels linguistic knowledge may help here but this is clearly a case where statistical knowledge should also play a part. For the NLP community issues of quantitative evaluation need to be addressed. Her comment:

> The biggest disadvantage of statistical models may be a lack of familiarity to those more comfortable with symbolic approaches.

and vice versa, the lack of linguistic knowledge by those from a statistical background, is certainly a very important obstacle to be overcome.

The chapter by Ramshaw and Marcus delves into the properties of Brill's transformation based learning specifically looking at its application to part of speech tagging, the application it has been most widely used for. The transformation-based learner uses frequency counts to find an optimum ordered list of rules to account for the training corpus starting from symbolic templates which determine the possible features to consider. The resultant symbolic rules have no attached probabilities. The chapter is an interesting comparison of the properties of transformation-based learning, HMMs and decision trees (with which transformation-based learning is more closely related). These properties include differences in the types of features that can be expressed and the remarkable ability of transformation-based learning to withstand over-training. A comparison of performance between the three approaches is relegated to future work. However, given this exploration of the different properties a comparison of performance should be made across a range of tasks and not just restricted to part-of-speech tagging.

The final chapter by Rosé and Waibel describes a speech to speech translation system where parser recovery is essential given all the additional difficulties that beset an NLP system when spontaneous speech is involved. The recovery involves fitting partial feature structures together using symbolic information to ensure the combination is meaningful and using mutual information between the fillers and slots to constrain the search. Further parameters are applied to limit the scope of what the parser will consider to ensure effort is spent only on likely hypotheses: this means that unlikely (but correct) representations can be missed. The main drawback of the system reported is the tendency to ask the user too many questions. The use of mutual information, whilst certainly being a popular statistic by virtue of its intuitive clarity, might merit some consideration because it breaks down when dealing with rare events (Dunning, 1993). Again, the issue of selecting the right statistical technique for the job is pertinent.

The book is a collection of quite heterogeneous papers in many different areas of NLP. The chapters provide a mixture of theoretical and practical perspectives, descriptions of general architectures and specific applications, and their focus varies from specific strategies, through evaluation methodologies to background overviews on the "symbolic" and "statistical" orientations. Perhaps because of the diversity of the topics covered there is no attempt to group the chapters into sections and the order appears rather arbitrary. The editors do however provide a useful introduction to each individual chapter.

The unifying theme to the eight chapters is the pursuit of the two goals, that symbolic and statistical approaches can and should be combined and how this can be performed. As Price points out, whilst researchers from the two camps may wish to utilise techniques from the other, lack of knowledge is a significant hurdle to be overcome. Comparisons of statistics such as that given by Daille, and of linguistic knowledge like that given by Hatzivassiloglou are very useful. As pointed out by Dunning (Dunning, 1993) many NLP researchers without a statistical background may be tempted to use statistics without ensuring the required assumptions are met. Baayen and Sproat (Baayen and Sproat, 1996) also demonstrate the pitfalls of crude handling of unseen events. These events can provide a substantial source for error because they are common place in natural language corpora. Likewise, the advantages of using linguistic knowledge will be amplified if the knowledge is well-founded.

A book of this size can only cover so much ground and there are many untouched areas and outstanding issues. The volume is slim but reasonably priced and is an interesting collection for readers wishing to dip into a variety of topics rather than being of particular merit in any one area.

References

Baayen, Harald and Sproat, Richard. 1996. Estimating Lexical Priors for Low-Frequency Morphologically Ambigous Forms. *Computational Linguistics* 22: 155-166

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19: 61-74

Gazdar, Gerald. 1996. Paradigm merger in natural language processing. In Milner,R. and Wand, I. (Eds.), *Computing Tomorrow: Future Research Directions in Computer Science* pp 88-109. Cambridge University Press. Cambridge, UK.

Grefenstette, Gregory. 1992. Use of Syntactic Context to produce Term Association Lists for Text Retrieval. In Belkin, N., Ingwersen,P. and Pejtersen,A. (Eds.), *15th*

Annual Internations ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 89-97 Copenhagen, Denmark.

Pereira,Fernando and Tishby, Nattali and Lee,Lillian. 1993. Distributional Clustering of English Words In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguists pp 183-190*

Resnik, Philip, 1993 Selection and Information: A Class-Based Approach to Lexical Relationships. *Ph.D. Thesis, University of Pennsylvania*