Machine Learning, , 1–5 (1999)

Guest Editors' Introduction: Machine Learning and Natural Language

CLAIRE CARDIE cardie@cs.cornell.edu
Department of Computer Science, Cornell University, Ithaca, NY 14853-7501

RAYMOND J. MOONEY mooney@cs.utexas.edu
Department of Computer Sciences, Taylor Hall 2.124, University of Texas, Austin, TX 78712-1188

The application of machine learning techniques to natural language processing (NLP) has increased dramatically in recent years under the name of "corpus-based," "statistical," or "empirical" methods. However, most of this research has been conducted outside the traditional machine learning research community. This special issue attempts to bridge this divide by assembling an interesting variety of recent research papers on various aspects of natural language learning – many from authors who do not generally publish in the traditional machine learning literature – and presenting them to the readers of Machine Learning.

In the last five to ten years there has been a dramatic shift in computational linguistics from manually constructing grammars and knowledge bases to partially or totally automating this process by using statistical learning methods trained on large annotated or unannotated natural language corpora. The success of statistical methods in speech recognition (Stolcke, 1997; Jelinek, 1998) has been particularly influential in motivating the application of similar methods to other aspects of natural language processing. There is now a variety of work on applying learning methods to almost all other aspects of language processing as well (Brill and Mooney, 1997), including morphological and syntactic analysis (Charniak, 1997), semantic disambiguation and interpretation (Ng and Zelle, 1997), discourse processing and information extraction (Cardie, 1997), and machine translation (Knight, 1997). Some concrete publication statistics clearly illustrate the extent of the revolution in natural language research. According to data recently collected by Hirschberg (1998), a full 63.5% of the papers in the Proceedings of the Annual Meeting of the Association for Computational Linguistics and 47.4% of the papers in the journal *Computational Linguistics* concerned corpus-based research in 1997. For comparison, 1983 was the last year in which there were no such papers and the percentages in 1990 were still only 12.8% and 15.4%.

Nevertheless, traditional machine learning research in artificial intelligence has had limited influence on recent research in computational linguistics. This is unfortunate since we believe that machine learning and empirical NLP have much to offer each other and that increased interaction and exchange of ideas would greatly benefit both areas. Most current learning research in NLP employs particular statistical techniques inspired by research in speech recognition, such as hidden Markov models (HMMs) and probabilistic context-free grammars (PCFGs). A variety of other learning methods including decision tree and rule induction, neural networks, instance-based methods, Bayesian network learning, inductive logic programming, explanation-based learning, and genetic algorithms can also be applied to natural language problems and can have significant advantages in particular applications. In addition to specific learning algorithms, a variety of general ideas from traditional machine learning such as active learning, boosting, reinforcement learning, constructive induction, learning with background knowledge, theory refinement, experimental evaluation methods, PAC learnability, etc., may also be usefully applied to natural language problems. In fact, a number of the papers in this collection illustrate the utility of several of these specific methods and ideas.

Within machine learning, there seems to be a growing dissatisfaction with the somewhat narrow focus on feature-vector classification problems represented by the standard data sets available from the UCI Repository. The computational linguistics community has assembled large data sets on a range of interesting NLP problems. Some of these problems can be reduced to a standard classification task by appropriately constructing features; however, others require using and/or producing complex data structures such as complete sentences or parse trees.[1]

Therefore, machine learning can provide natural language processing a range of alternative learning algorithms as well as additional general approaches and methodologies. On the other hand, natural language can provide machine learning with a variety of interesting, important, and challenging problems, frequently with particular characteristics such as a very large feature space and extremely sparse data. A less obvious potential contribution of statistical NLP to machine learning is the introduction of new learning methods. Several learning techniques in statistical NLP, such as the maximum entropy methods (Ratnaparkhi, this volume) and exponential models (Beeferman, Berger, & Lafferty, this volume), are not well studied in the machine learning literature. Perhaps these methods could also be productively applied to other machine learning problems. For example, the success of HMMs and PCFGs in language has already motivated interesting applications in molecular biology (Krogh, Brown, Mian, Sjolander, and Haussler, 1994; Abe and Mamitsuka, 1997).

As shown in Table 1, the papers in this issue tackle a broad spectrum of language-learning tasks using a variety of machine learning algorithms. Nearly half of the articles rely on algorithms initially developed within the machine learning community. Daelemans, van den Bosch, & Zavrel use instance-based (memory-based, case-based) learning (Stanfill and Waltz, 1986; Aha, Kibler, and Albert, 1991). Golding & Roth combine the perceptron-like Winnow algorithm (Littlestone, 1988) and weighted-majority voting (Littlestone and Warmuth, 1994), both of which originated in the computational learning theory community. Decision tree induction (Quinlan, 1986, 1992) and boosting (Freund and Schapire, 1996) are the primary methods employed by Haruno, Shirai, & Ooyama. Soderland presents a task-specific, general-to-specific rule-learning algorithm that is similar to existing top-down rule induction systems like FOIL (Quinlan, 1990). The remaining articles in the issue employ statistical methods that are more commonly encountered in the speech-processing and statistical NLP communities. Bikel, Schwartz, & Weischedel use HMMs. Dagan, Lee, & Pereira develop probabilistic word association models based on nearest-neighbor clustering techniques. Brent also presents a probabilistic model, but for word segmentation rather than word association. As noted in the paragraphs above, Ratnaparkhi uses maximum entropy modeling and Beeferman *et al.* employ exponential language modeling techniques.

Table 1. Summary of Papers in the Special Issue on Natural Language Learning

Paper	ML Technique(s)	NLP Task(s)
Beeferman *et al.*	exponential language modeling	text segmentation
Bikel *et al.*	Hidden Markov Models	named entity identification
Brent	probabilistic	word segmentation
Daelemans *et al.*	instance-based learning	grapheme-to-phoneme conversion, part-of-speech tagging, noun phrase chunking, prepositional phrase attachment
Dagan *et al.*	probabilistic; similarity-based learning	handling sparse data, pseudo word-sense disambiguation, language modeling
Golding & Roth	Winnow	context-sensitive spelling correction
Haruno *et al.*	decision trees; boosting	syntactic parsing
Ratnaparkhi	maximum entropy modeling	syntactic parsing
Soderland	rule-learning	concept extraction

In terms of performance tasks, the papers in this issue tackle a diverse collection of language-learning problems from low-level linguistic analysis through high-level NLP applications. At the sub-lexical level, Daelemans *et al.* address the problem of grapheme-to-phoneme (text-to-speech) conversion; Brent presents a cognitively plausible, unsupervised method for learning word boundaries in speech. Moving up to lexical tagging tasks, the Daelemans *et al.* article covers part-of-speech tagging and noun phrase chunking; Dagan *et al.* evaluate their clustering techniques on a pseudo word-sense disambiguation task. Algorithms for making high-level structural attachment decisions for sentence analysis are also represented in the special issue: Daelemens *et al.* apply instance-based techniques to the problem of prepositional phrase attachment; Haruno *et al.* and Ratnaparkhi present work on syntactic parsing of Japanese and English, respectively. Corpus-based approaches to all of the above problems typically encounter sparse data problems — it is difficult to determine the likelihood of many linguistic events because they occur infrequently or not at all in the training corpus. Dagan *et al.* present a general solution to this problem in their similarity-based clustering techniques. As a result, their work can be viewed as task-independent supporting technology.

A number of end-to-end NLP applications are also represented. Golding & Roth look at the problem of context-sensitive spelling correction. Two articles tackle problems that arise in building information extraction systems — NLP systems that scan an arbitrary input document for information on a particular topic (e.g., natural disasters) and produce as output a template that summarizes the major findings with respect to that topic (e.g., the time, date, and location of the disaster; the number of people injured; the amount of property damage). In particular, Soderland presents a technique for learning concept extraction patterns and Bikel *et al.* present a system that learns to accurately identify names, dates, times, and numerical expressions. Finally, Beeferman *et al.* address the problem of text segmentation, a problem traditionally explored in the field of information retrieval.

The order of presentation of the papers is in terms of overall task: underlying learning methods (Daelemans *et al.*; Dagan *et al.*); lexical tasks (Brent; Golding & Roth); syntactic analysis (Haruno *et al.*; Ratnaparkhi); and extracting text structure and content (Beeferman *at al.*; Bikel *et al.*; Soderland). Together, the papers show how the full spectrum of natural language understanding tasks can be successfully mapped to machine learning problems. We believe that the papers in this special issue provide a particularly good example of the potential advantages of interaction between learning and language research and hope it encourages further communication between the ML and NLP communities. As an additional source of papers on empirical, corpus-based approaches to natural language processing, we recommend that the reader consult the recent proceedings of the major conferences in the area such as the Annual Conference of the Association for Computational Linguistics and COLING (Boitet and Whitelock, 1998; Cohen and Wahlster, 1997), the Conference on Empirical Methods in NLP (Ide and Voutilainen, 1998; Cardie and Weischedel, 1997), and the Workshop on Very Large Corpora (Charniak, 1998; Zhou and Church, 1997).

Acknowledgments

We would like to sincerely thank all of the authors, the many anonymous reviewers, and Tom Dietterich as executive editor, without whom this special issue would not have been possible.

Notes

1. The Linguistic Data Consortium (`www.ldc.upenn.edu`) is one of the primary sources of data in NLP. Additional information is available from the Association of Computational Linguistics' special interest groups on Linguistic Data (`www.cs.jhu.edu/~ yarowsky/sigdat.html`) and on Natural Language Learning (`www.cs.unimaas.nl/signll/signll-home.html`).

References

Abe, N., & Mamitsuka, H. (1997). Predicting protein secondary structure using stochastic tree grammars. *Machine Learning*, *29*, 275–301.

Aha, D., Kibler, D., & Albert, M. (1991). Instance-Based Learning Algorithms. *Machine Learning*, *6*(1), 37–66.

Boitet, C., & Whitelock, P. (Eds.). (1998). *Proceedings of the 36th Annual Meeting of the ACL and COLING-98.* Morgan Kaufmann, San Francisco, CA.

Brill, E., & Mooney, R. J. (1997). An overview of empirical natural language processing. *AI Magazine*, *18*(4), 13–24.

Cardie, C. (1997). Empirical methods in information extraction. *AI Magazine*, *18*(4), 65–79.

Cardie, C., & Weischedel, R. (Eds.). (1997). *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing.* Association of Computational Linguistics.

Charniak, E. (1997). Statistical techniques for natural language parsing. *AI Magazine*, *18*(4), 33–43.

Charniak, E. (Ed.). (1998). *Proceedings of the Sixth Workshop on Very Large Corpora.* Association of Computational Linguistics Special Interest Group on Linguistic Data.

Cohen, P., & Wahlster, W. (Eds.). (1997). *Proceedings of the 35th Annual Meeting of the ACL and EACL-97.* Morgan Kaufmann, San Francisco, CA.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148–156. Morgan Kaufmann, San Francisco, CA.

Hirschberg, J. (1998). Every time I fire a linguist, my performance goes up, and other myths of the statistical natural language processing revolution. Invited talk, Fifteenth National Conference on Artificial Intelligence (AAAI-98).

Ide, N., & Voutilainen, A. (Eds.). (1998). *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing.* Association of Computational Linguistics.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, MA.

Knight, K. (1997). Automating knowledge acquisition for machine translation. *AI Magazine*, *18*(4), 81–96.

Krogh, A., Brown, M., Mian, S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, *235*, 1501–1531.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, *2*, 285–318.

Littlestone, N., & Warmuth, M. (1994). The weighted majority algorithm. *Information and Computaton*, *108*, 212–261.

Ng, H. T., & Zelle, J. (1997). Corpus-based approaches to semantic interpretation in natural language processing. *AI Magazine*, *18*(4), 45–64.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine Learning*, *5*, 239–266.

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA.

Stanfill, C., & Waltz, D. (1986). Toward Memory-based Reasoning. *Communications of the ACM*, *29*, 1213–1228.

Stolcke, A. (1997). Linguistic knowledge and empirical methods in speech recognition. *AI Magazine*, *18*(4), 25–31.

Zhou, J., & Church, K. (Eds.). (1997). *Proceedings of the Fifth Workshop on Very Large Corpora.* Association of Computational Linguistics Special Interest Group on Linguistic Data.