
Machine Learning, 60, 5–9, 2005

Guest Editors Introduction: Machine Learning in Speech and Language Technologies

PASCALE FUNG pascale@ee.ust.hk
Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Hong Kong

DAN ROTH danr@cs.uiuc.edu
Department of Computer Science, University of Illinois, Urbana, IL. 61801, USA

Machine learning techniques have long been the foundations of speech processing. Bayesian classification, decision trees, unsupervised clustering, the EM algorithm, maximum entropy, etc. are all part of existing speech recognition systems. The success of statistical speech recognition has led to the rise of statistical and empirical methods in natural language processing. Indeed, many of the machine learning techniques used in language processing, from statistical part-of-speech tagging to the noisy channel model for machine translation have roots in work conducted in the speech field.

However, advances in learning theory and algorithmic machine learning approaches in recent years have led to significant changes in the direction and emphasis of the statistical and learning centered research in natural language processing and made a mark on natural language and speech processing. Approaches such as memory based learning, a range of linear classifiers such as Boosting, SVMs and SNoW and others have been successfully applied to a broad range of natural language problems, and these now inspire new research in speech retrieval and recognition. We have seen an increasingly close collaboration between speech and language processing researchers in some of the shared tasks such as spontaneous speech recognition and understanding, voice data information extraction, and machine translation.

The purpose of this special issue was to invite speech and language researchers to communicate with each other, and with the machine learning community on the latest machine learning advances in their work. The call for papers was met with great enthusiasm from the speech and natural language community. Thirty six submissions were received; each paper was reviewed by at least three reviewers. Only ten papers were selected reflecting not only some of the best work on machine learning in the areas of natural language and spoken language processing but also what we view as a collection of papers that represent current trends in these areas of research both from the perspective of machine learning and from that of the speech and natural language applications perspective.

The papers in this special issue cover a broad range of topics in natural language and spoken language processing as well as in machine learning. From both perspectives the selection reflects the maturity of the field which has moved to address harder problems using more sophisticated techniques.

Early work in empirical methods in natural language processing was influenced by the success of statistical speech recognition and was dominated by relatively simple statistical methods. Many of the early works, from statistical part-of-speech tagging (Church, 1988; Church & Mercer, 1993) to the noisy channel model for machine translation, have roots in work conducted in the speech field.

Most of the early works can be viewed as based on generative probability models, which provide a principled way to the study of statistical classification. In these models, it is common to assume a generative model for the data, estimate its most likely parameters from training data and then use Bayes rule to obtain a classifier for this model. Naturally, estimating the most likely parameters involves making simplifying assumptions on the generating model.

Advances in Learning Theory and algorithmic Machine Learning in the last few years resulted in developing a better understanding of the relations between probabilistic models of classifications and discriminative models and had a significant effect on work in natural langauge processing. It has become clear (Roth, 1998, 1999) that probabilistic classifiers make use of the same representations as other existing classifiers, namely, a linear representation over some feature space. As a result, it is possible to keep using the same representation, but develop other ways of parameter estimation, driven directly by the eventual goal, to support better predictions (Collins, 2001). This understanding has led to a vast use of discriminative approaches such as memory based learning and a range of linear classifiers such as Boosting, SVMs, Winnow and Perceptron, all successfully applied to a broad range of natural language problems. It also inspired new research in speech retrieval and recognition.

The recent emphasis on discriminative methods applies not only to simple classification problems but also to machine learning work on more complex structured models.

Many problems in natural language processing involve assigning values to sets of variables where a complex and expressive structure can influence, or even dictate, what assignments are possible. Tasks such as labeling part-of-speech tags to the words of a sentence, many segmentation (e.g., shallow parsing) and parsing problems are key examples of such tasks.

Traditionally, solutions to these problems were generative, as represented by HMM models for sequence learning problems and shallow parsing. More recently, when people have started to use more discriminative methods for classification, structured problems were addressed by decoupling learning from the task of maintaining structured output. Only after estimators are learned for each local output variable are they used to produce global output consistent with the structural constraints. Discriminative HMM, conditional models (Punyakanok & Roth, 2001; McCallum, Freitag, & Pereira, 2000) and many dynamic programming based schemes used in the context of sequential predictions fall into this category.

Another class of solutions has been developed by realizing that even for complex models there is a way to incorporate dependencies among the variables into the learning process, and directly induce estimators to optimize a global performance measure (Roth, 1999; Lafferty, McCallum, & Pereira, 2001; Collins, 2002). Understanding the tradeoffs between different approaches to learning is currently an active area of research (Punyakanok et al., 2005).

Several of the papers in this special issue can be viewed from this perspective—they study discriminative models of learning in the context of structure learning.

The work of Pradhan et al., is concerned with an area of growing interest to the natural language processing community—that of shallow semantic parsing (semantic role labeling)—the process of assigning a *who did what to whom, when, where, why, how* etc. structure to plain text (Carreras & Màrquez, 2004; Kingsbury & Palmer, 2002). They view the problem as a phrase segmentation and identification process, and apply a process which decouples the learning of word-based classifiers (in this case, a support vector machine classifier identifying, for example, words that are inside a specific verb argument) from that of enforcing some simple sequential constraints; basically following a conditional model paradigm (Punyakanok & Roth, 2001; McCallum, Freitag, & Pereira, 2000).

Carreras, Marquez and Castro, on the other hand, also study a phrase segmentation problem, but develop an approach in which the learning algorithm is coupled with a level of inference that enforces some constraints among a number of classifiers. They learn by using a version of the perceptron learning algorithm in which feedback is driven by global, task level error.

Vast use of linear learning algorithms such as perceptron and its variations has been one of the characteristics of the move to discriminative models mentioned above. Shen and Joshi develop variants of perceptron in the context of *Ranking* tasks, which lie between classification and regression problems, and use them for parsing problems and machine translation.

The paper by Alshawi also makes use of variants of on-line linear learning algorithms such as perceptron and winnow in the multiclass setting this time, but develops versions of these that are more suited to the limited feedback setting, while sharing the efficiency advantages of the standard ones. Alshawi studies these algorithms in the context of spoken language applications that adapt from feedback rather than being trained in batch mode, such as that of utterance classification.

The great amount of work on discriminative learning algorithms, especially linear learning algorithms, brought more attention to work on features. When using linear learning algorithms it is necessary to learn over expressive features. In most cases, this is done explicitly by extracting features that are local conjunctions (n-grams) of words, part-of-speech tags or other pieces of information available in the input. For algorithms like support vector machines and perceptron, it is possible to generate features implicitly, via the notion of *kernels*.

In the last few years there has been interest in the natural language community in developing kernels that are appropriate for NLP applications (Collins & Duffy, 2002; Cumby & Roth, 2003). Although there have been several studies which considered the advantages and disadvantages of using explicit feature generation vs. implicit kernels (Cumby & Roth, 2003; Kudo & Matsumoto, 2003) the study of expressive kernels is very important for natural langauge applications.

Cortes and Mohri study kernel methods that can deal with variable length sequences or, more generally, distributions given by weighted automata, and are useful in applications such as spoken-dialog classification.

In addition to variants of perceptron, winnow and support-vector-machine that were used in several papers discussed above, another class of linear models that is popular in natural language applications is that of maximum entropy models. In maximum entropy based classifiers a conditional distribution is estimated as a log-linear model over a collection of

features. Garcia-Varea and Casacuberta make use of these features in order to add contextual information to statistical lexicon and improve the performance of the statistical translation systems. Kazama and Tsujii extend the standard maximum entropy model in an attempt to alleviate data sparseness in parameter estimation. This is done by developing a form of regularization—allowing the use of box-type inequality constraints, where equality can be violated up to certain predefined levels.

While all papers discussed so far developed machine learning techniques in the context of natural language applications, the last three papers make more direct use of statistical techniques.

Two papers, by Emami and Jelinek and by Wang et al. address the long standing challenge of incorporating higher level information other than lexical n-grams to language models. Incorporating structural information in language models proved to be beneficial to machine translation tasks. Statistical n-gram models have enjoyed widespread use mostly because they are computationally efficient with relatively good performance. However, n-grams do not capture the long distance dependency between words in a sentence, nor do they encapsulate any syntactic structure. A huge amount of training data is required for a robust n-gram model. Above-mentioned papers in this issue attempt to alleviate these problems using different learning models. Emami and Jelinek propose a neural probabilistic model in a syntactic based language model. In their approach, each token is associated with a feature vector representing its history, the neural network then estimates the next probable word given a concatenation of input feature vectors. According to them, the neural network can handle larger vocabularies with longer contexts then n-gram models. Their experimental results on large test sets have shown great promise for using neural net based syntactic language models. The approach of Wang et al. also incorporates various higher level information about words, such as syntax and semantics, in a unified probabilistic framework. They propose the latent maximum entropy principle to estimate the hidden hierarchical structure of natural language without requiring explicit parse tree or semantic labels, thus alleviating the sparse data problem.

Finally, the paper by Turney and Littman makes use of relatively well-known methods, based on the vector space model of information retrieval, to address new semantic level processing questions—verbal analogy questions of the type found in SAT exams.

Overall, this special issue reflects both a variety of machine learning and statistical methods that are at the center of the research in natural language and spoken language processing, and a range of applications studied in these areas. Papers in this issue provide a particularly good example for works that advance the research both from the machine learning and from the language processing perspective, and in this way exemplify the potential advantages of interaction between learning and language research. We hope that it will encourage further communication and interaction between these research communities.

Acknowledgments

We would like to sincerely thank all of the authors and the many reviewers that contributed to this special issue. Special thanks go the Machine Learning Journal editorial staff.

The work of Pascale Fung was partly supported by CERG #HKUST6206/03E of the Research Grants Council of the Hong Kong government.

The work of Dan Roth was supported by NSF grants ITR-IIS-0085836, ITR-IIS-0085980 and IIS-9984168 and an ONR MURI Award.

References

Carreras, X., & Màrquez, L. (2004). Introduction to the CoNLL-2004 shared tasks: Semantic role labeling. In *CoNLL: Conference on Natural Language Learning*.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *ANLP: Conference on Applied Natural Language Processing*.

Church, K. W., & Mercer, R. L. (1993). Introduction to the special issue in computational linguistics using large corpora. *Computational Linguistics, 19:1*, 1–24.

Collins, M. (2001). Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In *Proc. of the International Workshop on Parsing Technologies*.

Collins, M. (2002). Discriminative training methods for Hidden Markov models: Theory and experiments with perceptron algorithms. In *EMNLP: Conference on Emprirical Methods in Natural Language Processing*.

Collins, M., & Duffy, N. (2002). New ranking algorithms for parsing and taggins: Kernels over discrete structures, and the voted perceptron. In *Proc. of the Annual Meeting of the ACL.*

Cumby, C., & Roth, D. (2003). On Kernel methods for relational learning. In *Proc. of the International Conference on Machine Learning*.

Kingsbury, P., & Palmer, M. (2002). From Treebank to PropBank. In *Proc. of LREC*.

Kudo, T., & Matsumoto, Y. (2003). Fast methods for Kernel-based text analysis. In *Proc. of the Annual Meeting of the ACL.*

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the Eighteenth International Conference on Machine Learning* (pp. 282–289).

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. In *Proc. of the International Conference on Machine Learning*. Stanford, CA.

Punyakanok, V., & Roth, D. (2001). The use of classifiers in sequential inference. In *Advances in Neural Information Processing Systems 13*.

Punyakanok, V., Roth, D., Yih, W., & Zimak, D. (2005). Learning and inference over constrained output. In *Proc. of the International Joint Conference on Artificial Intelligence*.

Roth, D. (1998). Learning to resolve natural language ambiguities: A unified approach. In *Proc. of Ameirican Association of Artificial Intelligence* (pp. 806–813).

Roth, D. (1999). Learning in natural language. In *Proc. of the International Joint Conference on Artificial Intelligence* (pp. 898–904).