Symposium on Machine Learning in Speech and Language Processing

June 27, 2011
Bellevue, Washington, USA

Call for Participation

The goal of the symposium is to foster communication and collaboration between researchers in these synergistic areas, taking advantage of the nearby locations of ACL-HLT 2011 and ICML 2011. It will bring together members of the Association for Computational Linguistics, the International Speech Communication Association, and the International Machine Learning Society (Machine Learning Special Interest Group of ISCA).

Topics

The workshop will feature a series of invited talks and general submissions. Submissions focusing on novel research are solicited and we especially encourage position and review papers addressing topics that are relevant both to Speech, Machine Learning and NLP. These areas include but are not limited to the use of: SVMs, log-linear models, neural networks, kernel methods, discriminative transforms, large margin training, discriminative training, active, semi-supervised and unsupervised training, structured prediction, Bayesian modeling, deep learning , and sparse representations. Application areas include natural language processing, speech recognition, language modeling, and speaker verification.

Paper Submission

Prospective authors are invited to submit papers written in English via the "Submissions" link to the left. Each paper will be reviewed by at least two reviewers, and each accepted paper must have at least one registered author.

Invited Speakers

Yoshua Bengio, Jeff Bilmes, Ming-Wei Chang, Stanley Chen, Jason Eisner, Eduard Hovy, Sanjoy Dasgupta, Mark Hasegawa-Johnson, David McAllester, George Saon, Lawrence Saul, and Mark Steedman.

Organizing Committee

Hal Daume III	University of Maryland
Joseph Keshet	TTI-Chicago
Dan Roth	UIUC
Geoffrey Zweig	Microsoft

Scientific Program Committee

Jeff Bilmes University of Washington
Brian Kingsbury IBM
Karen Livescu TTI-Chicago

Symposium on Machine Learning in Speech and Language Processing

June 27, 2011
Bellevue, Washington, USA

Program

8:00-8:15	**Welcome**	
8:15-8:45	Sanjoy Dasgupta	Recent Advances in Active Learning
8:45-9:15	Lawrence Saul	Online learning of large margin hidden Markov models for automatic speech recognition
9:15-9:45	Eduard Hovy	On the Role of Machine Learning in NLP
9:45-10:15	George Saon	Bayesian Sensing Hidden Markov Models for Speech Recognition
10:15-10:30	**Break**	
10:30-11:00	Jason Eisner	*A Non-Parametric Bayesian Approach to Inflectional Morphology*
11:00-11:30	Mark Hasegawa-Johnson	Unlabeled Data and Other Marginals
11:30-12:00	Ming-Wei Chang	Structured Prediction with Indirect Supervision
12:00-1:30	**Lunch**	
1:30-3:00	**Poster Session**	
	Robert Moore and John DeNero	L_1 and L_2 Regularization for Multiclass Hinge Loss Models
	Shirin Badiezadegan and Richard Rose	A Comparison of Performance Monitoring Approaches to Fusing Spectrogram Channels in Speech Recognition
	Meng Sun and Hugo Van hamme	A Two-Layer Non-negative Matrix Factorization Model for Vocabulary Discovery
	Deryle Lonsdale and Carl Christensen	Automating the scoring of elicited imitation tests
	Rushin Shah, Bo Lin, Kevin Dela Rosa, Anatole Gershman and Robert Frederking	Improving Cross-document Co-reference with Semi-supervised Information Extraction Models

	David Chen and Raymond Mooney	Panning for Gold: Finding Relevant Semantic Content for Grounded Language Learning
	Hynek Hermansky, Nima Mesgarani and Samuel Thomas	Performance Monitoring for Robustness in Automatic Recognition of Speech
3:00-3:30	Jeff Bilmes	Applications of Submodular Functions in Speech and NLP
3:30-4:00	David McAllester	Generalization Bounds and Consistency for Latent-Structural Probit and Ramp Loss
4:00-4:30	Mark Steedman	Some Open Problems in Machine Learning for NLP
4:30-4:45	**Break**	
4:45-5:15	Stanley Chen	Performance Prediction and Shrinking Language Models
5:15-5:45	Yoshua Bengio	*On Learning Distributed Representations of Semantics*
7:00	**Invited Speakers Dinner at nearby restaurant**	

Invited Speakers

Recent Advances in Active Learning
Sanjoy Dasgupta

A key requirement for being able to learn a good classifier is having enough labeled data. In many situations, however, unlabeled data is easily available but labels are expensive to come by. In the active learning scenario, each label has a non-negligible cost, and the goal, starting with a large pool of unlabeled data, is to adaptively decide which points to label, so that a good classifier is obtained at low cost.

Many active learning strategies run into severe problems with sampling bias; the theory has therefore focused on how to correctly manage this bias while attaining good label complexity. I will summarize recent work in the machine learning community that achieves this goal through algorithms that are simple and practical enough to be used in large-scale applications.

Online learning of large margin hidden Markov models for automatic speech recognition
Lawrence Saul

We explore the use of sequential, mistake-driven updates for online learning and acoustic feature adaptation in large margin hidden Markov models. The updates are applied to the parameters of acoustic models after the decoding of individual training utterances. For large margin training, the updates attempt to separate the log-likelihoods of correct and incorrect transcriptions by an amount proportional to their Hamming distance. For acoustic feature adaptation, the updates attempt to improve recognition by linearly transforming the features

computed by the front end. We evaluate acoustic models trained in this way on the TIMIT speech database. We find that online updates for large margin training not only converge faster than analogous batch optimizations, but also yield lower phone error rates than approaches that do not attempt to enforce a large margin. Finally, experimenting with different schemes for initialization and parameter-tying, we find that acoustic feature adaptation leads to further improvements beyond the already significant gains achieved by large margin training.

On the Role of Machine Learning in NLP

Eduard Hovy

Almost all of NLP consists of the following three kinds of tasks: transforming information from one representation into another, identifying within a larger collection a fragment or subset obeying given desiderata, and assigning an appropriate label to a given exemplar. These tasks used to be performed by algorithms using manually crafted rules. The power, and the promise, of Machine Learning is to perform much of this work automatically. Over the past 20 years, NLP researchers have explored many kinds of learning algorithms on existing and/or easily created corpora for a wide range of tasks and phenomena. By using existing corpora or cleverly leveraging resources, they have avoided the difficult tasks of designing representations and building training data. But nothing is free forever. To make headway achieve increasingly high performance in various NLP tasks, we are going to need some deeper thinking about the phenomena themselves, about suitable representations for them, and about corpora that illustrate their complexity and scale. The core problem is that Machine Learning focuses on only half of the problem: it has nothing to say about the nature of the phenomena being addressed, and is not, ultimately, very useful for arriving at a deeper understanding of how language works. The best way forward for NLP, I believe, is to recognize that we need (at least) two kinds of researchers: the NLP linguists and the NLP engineers. In this talk I outline the problem and suggest ways in which ML researchers (who mostly are NLP engineers) can facilitate the work of NLP linguists. This effort will be repaid handsomely, since there remain many challenging problems in NLP, ready to be addressed once the linguistic perspectives and representations have been addressed.

Bayesian Sensing Hidden Markov Models for Speech Recognition

George Saon

We introduce Bayesian sensing hidden Markov models to represent speech data based on a set of state-dependent basis vectors. By incorporating the prior density of sensing weights, the relevance of a feature vector to different bases is determined by the corresponding precision parameters. The model parameters, consisting of the basis vectors, the precision matrices of the sensing weights and the precision matrices of the reconstruction errors, are jointly estimated by maximizing the likelihood function, which is marginalized over the weights. We derive recursive solutions for the three parameters, which are expressed via maximum a posteriori estimates of the sensing weights.

This model was fielded in the latest DARPA GALE Arabic Broadcast News transcription evaluation and has shown gains on the evaluation data over state-of-the-art discriminatively trained HMMs with conventional Gaussian mixture models.

A Non-Parametric Bayesian Approach to Inflectional Morphology

Jason Eisner

We learn how the words of a language are inflected, given a plain text corpus plus a small

supervised set of known paradigms. The approach is principled, simply performing empirical Bayesian inference under a straightforward generative model that explicitly describes the generation of

1. The grammar and subregularities of the language (via many finite-state transducers coordinated in a Markov Random Field).
2. The infinite inventory of types and their inflectional paradigms (via a Dirichlet Process Mixture Model based on the above grammar).
3. The corpus of tokens (by sampling inflected words from the above inventory).

Our inference algorithm cleanly integrates several techniques that handle the different levels of the model: classical dynamic programming operations on the finite-state transducers, loopy belief propagation in the Markov Random Field, and MCMC and MCEM for the non-parametric Dirichlet Process Mixture Model.

We will build up the various components of the model in turn, showing experimental results along the way for several intermediate tasks such as lemmatization, transliteration, and inflection. Finally, we show that modeling paradigms jointly with the Markov Random Field, and learning from unannotated text corpora via the non-parametric model, significantly improves the quality of predicted word inflections.

This is joint work with Markus Dreyer.

Unlabeled Data and Other Marginals

Mark Hasegawa-Johnson

Machine learning minimizes bounds on E[h] computed over an unknown distribution p(x,y). Unlabeled data describe p(x), while scientific prior knowledge can describe p(y). This talk will discuss the use of unlabeled data to compute p(x), and of articulatory phonology to compute p(y), for acoustic modeling and pronunciation modeling in automatic speech recognition. We will demonstrate that, if either p(x) or p(y) is known, it's possible to substantially reduce the VC dimension of the function space, thereby substantially reducing the expected risk of the classifier. As a speculative example, we will show that p(y) can be improved (relative to usual ASR methods) using finite state machines based on articulatory phonology, and preliminary results will be reviewed. As a more fully developed example, we will show that the VC dimension of a maximum mutual information (MMI) speech recognizer can be bounded by the conditional entropy of y given x; the resulting training criterion is MMI over labeled data, minus conditional label entropy of unlabeled data. Algorithms and experimental results will be provided for the cases of isolated phone recognition, and of retraining using N-best lists.

This is joint work with Jui-Ting Huang and Xiaodan Zhuang.

Structured Prediction with Indirect Supervision

Ming-Wei Chang

Machine learning (ML) has already made significant impacts on our daily life. From hand-written digit recognition, spam filtering to ranking search results, machine learning techniques help us build intelligent systems more easily and make computers seem smarter. Nevertheless, current ML techniques support limited set of supervision protocols, making it difficult to transfer human knowledge to machines efficiently without labeling examples explicitly. However, structured tasks, which involve many interdependent decisions for a given example, are expensive to label. Given that many important tasks in natural language processing and information extraction are structured tasks, it is important to develop learning

frameworks that can use knowledge resources and other sources of indirect supervision in addition to labeled examples for the current task.

Applications of Submodular Functions in Speech and NLP

Jeff Bilmes

Submodular functions have a long history in economics, game theory, combinatorial optimization, electrical networks, and operations research. A submodular function operates on subsets of a finite ground set, and has certain properties that make optimization either tractable or approximable whereas otherwise neither would be possible. In this talk, after briefly reviewing relevant properties, we will discuss several applications of submodularity in speech and natural language processing. First, we will see how submodular functions can be used to address the problem of finding a subset of a large speech corpus that is useful for accurately and rapidly prototyping novel and computationally expensive speech recognition architectures. The principle partition of a submodular function allows rapid exploration of the tradeoff between limiting the number and quality of types vs. the number and quality of tokens in such a corpus. Secondly, we present a class of submodular functions useful for document summarization tasks, each combining fidelity and diversity terms. We show better than existing state-of-art results in both generic and query-focused document summarization on the DUC 2004-2007 evaluations. We also show that some well-known methods for document summarization are in fact submodular. Lastly, given time, we may present new methods for active semi-supervised learning on submodular functions.

Joint work with Hui Lin and Andrew Guillory.

Generalization Bounds and Consistency for Latent-Structural Probit and Ramp Loss

David McAllester

Linear predictors are scale-insensitive --- the prediction does not change when the weight vector defining the predictor is scaled up or down. This implies that direct regularization of the performance of a linear predictor with a scale sensitive regularizer (such as a norm of the weight vector) is meaningless. Linear predictors are typically learned by introducing a scale-sensitive surrogate loss function such as the hinge loss of an SVM. However, no convex surrogate loss function can be consistent in general --- in finite dimension SVMs are not consistent. Here we generalize probit loss and ramp loss to the latent-structural setting and show that both of these loss functions are consistent in arbitrary dimension for an arbitrary bounded task loss. Empirical experience with probit loss and ramp loss will be briefly discussed.

Some Open Problems in Machine Learning for NLP

Mark Steedman

Natural language processing is obstructed by two problems: that of ambiguity, and that of skewed distributions. Together they engender acute sparsity of data for supervised learning, both of grammars and parsing models.

The paper expresses some pessimism about the prospects for getting around this problem using unsupervised methods, and considers the prospects for finding naturally labeled datasets to extend supervised methods.

Performance Prediction and Shrinking Language Models
Stanley Chen

In this talk, we present a simple empirical law that vastly outperforms the Akaike and Bayesian Information Criterions at predicting the test set likelihood of an exponential language model. We discuss under what conditions this relationship holds; how it can be used to improve the design of language models; and whether these ideas can be applied to other types of statistical models as well. Specifically, we show how this relationship led to the design of "Model M", a class-based language model that outperforms all previous models of this type.

On Learning Distributed Representations of Semantics
Yoshua Bengio

Machine learning algorithms try to characterize configurations of variables that are plausible (somehow similar to those seen in the training set, and predictive of those that could be seen in a test set) so as to be able to answer questions about these configurations. A general approach towards this goal is to learn *representations* of these configurations that help to generalize to new configurations. We expose the statistical advantages of representations that are *distributed* and *deep* (at multiple levels of representation) and survey some of the advances in such feature learning algorithms, along with some of our recent work in this area, for natural language processing and pattern recognition. In particular, we highlight our effort towards modeling semantics beyond single-word embeddings, to capture relations between concepts and produce models of 2-argument relations such as (subject, verb, object) seen as (argument1, relation, argument2) that can be used to answer questions, disambiguate text, and learn from free text and knowledge bases in the same representational space.

Posters

L_1 and L_2 Regularization for Multiclass Hinge Loss Models
Robert Moore and John DeNero

This paper investigates the relationship between the loss function, the type of regularization, and the resulting model sparsity of discriminatively-trained multiclass linear models. The effects on sparsity of optimizing log loss are straightforward: L_2 regularization produces very dense models while L_1 regularization produces much sparser models. However, optimizing hinge loss yields more nuanced behavior. We give experimental evidence and theoretical arguments that, for a class of problems that arises frequently in natural-language processing, both L_1- and L_2-regularized hinge loss lead to sparser models than L_2-regularized log loss, but less sparse models than L_1-regularized log loss. Furthermore, we give evidence and arguments that for models with only indicator features, there is a critical threshold on the weight of the regularizer below which L_1- and L_2-regularized hinge loss tends to produce models of similar sparsity.

A Comparison of Performance Monitoring Approaches to Fusing Spectrogram Channels in Speech Recognition
Shirin Badiezadegan and Richard Rose

Implementations of two performance monitoring approaches to feature channel integration in robust automatic speech recognition are presented. These approaches combine multiple

feature channels, where the first one uses an open loop entropy-based criterion and the second one, motivated by psychophysical evidence in human speech perception, employs a closed loop criterion relating to the overall performance of the system. The multiple feature channels correspond to an ensemble of reconstructed spectrograms generated by applying multiresolution discrete wavelet transform analysis-synthesis filter-banks to corrupted speech spectrograms. The spectrograms associated with these feature channels differ in the degree to which information has been suppressed in multiple scales and frequency bands. The performance of these approaches is evaluated in the Aurora 3 speech in noise task domain.

A Two-Layer Non-negative Matrix Factorization Model for Vocabulary Discovery

Meng Sun and Hugo Van hamme

A two-layer NMF model is proposed for vocabulary discovery. The model first extracts low-level vocabulary patterns based on a histogram of co-occurrences of Gaussians. Then latent units are discovered by spectral embedding of Gaussians at layer-1. Layer-2 discovers vocabulary patterns based on the histogram of co-occurrences of the latent units. Improvements in unordered word error rates are observed from the low-level representation to the two-layer model on the Aurora2/Clean database. The relation between the latent units and the states of an HMM is discussed.

Automating the scoring of elicited imitation tests

Deryle Lonsdale and Carl Christensen

This paper explores the role of machine learning in automating the scoring for one kind of spoken language test: elicited imitation (EI). After sketching the background and rationale for EI testing, we give a brief overview of EI test results that we have collected. To date, the administration and scoring of these tests have been done sequentially and the scoring latency has not been critically important; our goal now is to automate the test. We show how this implies the need for an adaptive capability at run time, and motivate the need for machine learning in the creation of this kind of test. We discuss our sizable store of data from prior EI test administrations. Then we show various experiments that illustrate how this prior information is useful in predicting student performance. We present simulations designed to foreshadow how well the system will be able to adapt on-the-fly to student responses. Finally, we draw conclusions and mention possible future work.

Improving Cross-document Co-reference with Semi-supervised Information Extraction Modelsi

Rushin Shah, Bo Lin, Kevin Dela Rosa, Anatole Gershman and Robert Frederking

In this paper, we consider the problem of cross-document co-reference (CDC). Existing approaches tend treat CDC as an information retrieval based problem and used features such as TF-IDF cosine similarity to cluster documents and/or co-reference chains. We augmented these features with features based on biographical attributes, such as occupation, nationality, gender, etc., obtained by using semi-supervised attribute extraction models. Our results suggest that the addition of these features boosts the performance of our CDC system considerably. The extraction of such specific attributes allows us to use features, such as semantic similarity, mutual information and approximate name similarity which have not been used so far for CDC with traditional bag-of-words models. Our system achieves F scores of 0.82 and 0.81 on the WePS-1 and WePS-2 datasets, which rival the best reported scores for this problem.

Panning for Gold: Finding Relevant Semantic Content for Grounded Language Learning
David Chen and Raymond Mooney

One of the key challenges in grounded language acquisition is resolving the intentions of the expressions. Typically the task involves identifying a subset of records from a list of candidates as the correct meaning of a sentence. While most current work assume complete or partial independence between the records, we examine a scenario in which they are strongly related. By representing the set of potential meanings as a graph, we explicitly encode the relationships between the candidate meanings. We introduce a refinement algorithm that first learns a lexicon which is then used to remove parts of the graphs that are irrelevant. Experiments in a navigation domain shows that the algorithm successfully recovered over three quarters of the correct semantic content.

Performance Monitoring for Robustness in Automatic Recognition of Speechi
Hynek Hermansky, Nima Mesgarani and Samuel Thomas

A new method to deal with an unexpected harmful variability (noise) in speech during the operation of the system is reviewed. The fundamental idea is to derive in the training phase statistics of the system output for the data on which the system was trained and adaptively modify the system based on statistics derived during the operation. Multiple processing channels are formed by extracting different spectral and temporal modulation components from the speech signal. Information in each channel is used to estimate posterior probabilities of speech sounds (posteriogram) in each channel, and these estimates are fused to derive the final posteriogram. The autocorrelation matrix of the final posteriogram is adopted as the measure that summarizes the system performance. Initial setup of the fusion module is found by cross-correlating the probability estimates with phoneme labels on training data. During an operation, the matrix derived on the training data serves as the desirable target and the fusion module is modified to optimize the system performance. Results on phoneme recognition from noisy speech indicate the effectiveness of the method.