Reducing Annotation Effort using Generalized Expectation Criteria

Gregory Druck, Gideon Mann, Andrew McCallum
gdruck@cs.umass.edu - University of Massachusetts, Amherst, MA 01003
gideon.mann@gmail.com - Google, Inc., 76 9th Ave., New York, NY 10011
mccallum@cs.umass.edu -University of Massachusetts, Amherst, MA 01003

Technical Report UM-CS-2007-62
DRAFT

November 30, 2007

Abstract

Generalized expectation (GE) criteria [McCallum *et al.*, 2007] are terms in objective functions that assign scores to values of model expectations. In this paper we introduce GE-FL, a method that uses GE to train a probabilistic model using associations between input features and classes rather than complete labeled instances. Specifically, here the expectations are model predicted class distributions on unlabeled instances that contain selected input features. The score function is the KL divergence from reference distributions estimated using feature-class associations. We show that a multinomial logistic regression model trained with GE-FL outperforms several baseline methods that use feature-class associations. Next, we compare with a method that incorporates feature-class associations into Boosting [Schapire *et al.*, 2002] and find that it requires 400 labeled instances to attain the same accuracy as GE-FL, which uses no labeled instances. In human annotation experiments, we show that labeling features is on average 3.7 times faster than labeling documents, a result that supports similar findings in previous work [Raghavan *et al.*, 2006]. Additionally, using GE-FL provides a 1.0% absolute improvement in final accuracy over semi-supervised training with labeled documents. The accuracy difference is often much more pronounced with only a few minutes of annotation, where we see absolute accuracy improvements as high as 40%.

1 Introduction

Supervised learning requires labeled instances, which are often costly to obtain. Semi-supervised learning methods are an appealing solution for reducing labeling effort. However, despite the recent increase in semi-supervised learning research, real applications of semi-supervised learning remain rare. Reasons for this may include the time and space complexity and reliance on sensitive hyperparameters of semi-supervised methods. Additionally, many methods make strong assumptions about the data that may hold in small, synthetic data sets, but tend to be violated in real-world data.

Instead, we want a simple, robust method that can help reduce annotation effort. We argue that even in the absence of labeled instances, one often possesses some prior knowledge about the task. In this paper, we propose a new semi-supervised learning method that incorporates one particular type of prior knowledge: associations betweens features and classes, into training. This is accomplished using generalized expectation (GE) criteria.

A GE criterion [McCallum *et al.*, 2007] assigns scores to values of a model expectation. In this paper, we consider score functions that can be described as a distance between a model expectation and a reference expectation.

GE is similar to the method of moments, but allows us to express arbitrary scalar preferences on expectations of arbitrary functions, rather than requiring equality between sample and model moments. We also note three important differences from traditional training objective functions for factor graphs. First, there need not be a one-to-one relationship between GE terms and model factors. For example, GE allows expectations on sets of variables that form a subset of model factors, or on sets of variables larger than model factors. Next, model expectations in different GE terms can be conditioned on different data sets. Finally, the reference expectation (or more generally, score function) can come from any source, including other tasks or human prior knowledge.

Mann and McCallum [2007] incorporate prior information about the class distribution into training using a special case of GE called label regularization. In this method, the score function is the KL-divergence between model predicted class distributions on unlabeled data and class priors. Label regularization outperforms other standard semi-supervised methods including expectation maximization, entropy regularization [Grandvalet and Bengio, 2004], and a representative graph method [Bengio *et al.*, 2006]. These performance gains are observed even when the other methods have access to the class prior. Importantly, this method can scale to millions of unlabeled instances and requires little parameter tuning. Unlike many other discriminative semi-supervised learning methods, there is no assumption of low-density regions between class boundaries.

In this paper, we extend the method of Mann and McCallum [2007] to incorporate prior knowledge not merely about class prior distributions, but about input features. We refer to this method as GE-FL: GE using feature labels. We say that a feature and a class are *associated* if the feature is a strong indicator of the class. For example, in the **baseball vs. hockey** text classification task we know that there is an association between the word feature *pitcher* and the class *baseball*, but not between the word feature *pitcher* and the class *hockey*. Note that a single feature may have multiple associated classes. We use the terms feature labels and feature-class associations interchangeably.

The objective function is composed of multiple GE terms that score the model's predicted class distribution for instances that contain selected input features. The score function encourages these expectations to match reference distributions that are estimated using feature-class associations. The ability to leverage feature-class associations using GE-FL suggests an alternate modality of supervision in which instead of labeling instances as to their class, an annotator labels features with the classes with which they are associated.

In experiments, we show that a multinomial logistic regression model trained using GE-FL outperforms several baseline methods that use feature-class associations. Next, we compare with a method that incorporates feature-class associations into Boosting [Schapire *et al.*, 2002] and see that it requires 400 labeled instances to attain the same accuracy as GE-FL, which uses no labeled instances. In human annotation experiments, we show that using GE-FL instead of semi-supervised training with labeled documents can take less time and provide a more accurate model. For example, after only one minute of annotation time, we can achieve 80% accuracy on the **ibm vs. mac** text classification problem using GE-FL, whereas even after ten minutes labeling documents the accuracy with entropy regularization [Grandvalet and Bengio, 2004] is 77%. More generally, labeling features is on average 3.7 times faster than labeling documents, a result that supports similar findings in previous work [Raghavan *et al.*, 2006]. Additionally, using GE-FL provides a mean 1.0% absolute improvement in final accuracy. The accuracy difference is often much more pronounced with only a few minutes of annotation, where we show absolute accuracy improvements as high as 40%.

2 Related Work

Semi-supervised learning methods use both labeled and unlabeled data during training. Zhu [2005] provides a thorough survey of work in semi-supervised learning. We highlight a few key methods here. The Expectation Maximization (EM) algorithm [Dempster *et al.*, 1977] applies naturally to the problem of "missing labels". However, EM can fail when the generative modeling assumptions are violated, and often performs worse than supervised training [Cozman and Cohen, 2006]. Transductive support vector machines [Joachims, 1999] and entropy regularization [Grandvalet and Bengio, 2004] are semi-supervised learning algorithms for

discriminative models that attempt to place the decision boundary in sparse regions of the input space. These methods can fail when classes overlap and often require extensive tuning of hyperparameters. There is also work in nonparametric semi-supervised methods, in particular graph-based methods [Zhu and Ghahramani, 2002; Zhu *et al.*, 2003] [Zhu and Ghahramani, 2002; Zhu *et al.*, 2003] which propagate labels from labeled to unlabeled instances, after projecting the data onto a low-dimensional manifold. These methods typically require a metric for computing the distance between instances.

In this paper, we aim to leverage prior knowledge during training. Jin and Liu [2005] provide an iterative algorithm for incorporating prior knowledge about class distributions into supervised discriminative training. In the first step of each iteration, the discriminative model is trained using both training and test data. The classes for the test instances are unavailable, so Jin and Liu use per-instance class distribution estimates during training. In the second step, the discriminative model parameters are held fixed and the per-instance class distributions are re-estimated. The distributions are chosen to minimize the divergence from the model predictions subject to the constraint that the overall predicted class distribution matches the prior distribution. This algorithm helps especially in cases when the class distribution of the training data is not representative of the true class distribution. This idea is related to label regularization [Mann and McCallum, 2007], but the technical details are quite different.

Schapire, Rochery, and Gupta [2002] and Wu and Srihari [2004] leverage prior information about features during training. Schapire, Rochery, and Gupta modify AdaBoost to choose weak learners that booth fit the labeled training data and fit a model estimated using human provided feature-class associations. In contrast, we do not require complete labeled instances to train a model with GE-FL. We compare with this approach in Section 5.2. Haghighi and Klein [2006] use prototypes, essentially the same type of feature-class associations we use here, to learn log-linear models for structured output spaces. The prototypes are used to hypothesize additional "soft" prototypes for features that are syntactically similar. All prototypes are then used as features during maximum likelihood training. In contrast, here we encourage the model to match its predictions on unlabeled data with reference distributions estimated from prototypes. In ongoing work we are also applying GE to models for structured output spaces.

Chang, Ratinov, and Roth [2007] propose an EM-like algorithm that incorporates prior constraints into semi-supervised training of structured output models. In the E-step, the inference procedure produces an N-best list of outputs according to the output's score under the model and penalties for violated constraints. In the M-step, the N-best list is used to re-estimate the model parameters.

Graça, Ganchev, and Taskar [2008] incorporate constraints into the EM algorithm in a more principled way. The E-step of EM can be interpreted as finding the distribution in some set that minimizes the KL divergence from the model prediction. Typically this set includes the model distribution, but Graça, Ganchev, and Taskar instead restrict the set to only include distributions that match some constraints. This can be interpreted geometrically as the information projection of the model expected distribution onto the space spanned by the constraints. The solution to this projection is a unique, exponential family distribution. In the M-step, the model parameters are re-estimated using the projected distribution. This method is related to GE, but there are a few differences. First, the constraints in this method are per-instance, whereas in this paper we use global constraints. Next, Graça et al. use a generative model that is trained with a modified EM algorithm, whereas here we use direct maximization and a discriminative model. Finally, Graça et al. put constraints only on the output variables, whereas our constraints additionally consider input variables.

Active learning is a related problem in which the learner can choose the particular instances to be labeled. In pool-based active learning [Cohn *et al.*, 1994], the learner has access to a set of unlabeled instances, and can choose the instance that has the highest expected utility according to some metric. A standard pool-based active learning method is uncertainty sampling [Lewis and Catlett, 1994], in which the instance chosen is the one for which the model predictions are most uncertain. Although in theory this method is problematic because it ignores the distribution over instances [Freund *et al.*, 1997], in practice it is easy to implement, and often works well. We compare against uncertainty sampling for choosing instances in our human labeling experiments.

Raghavan, Madani, and Jones [2006] interleave feedback on features into uncertainty sampling, and show that such feedback can significantly accelerate active learning. Experiments also demonstrate that humans

can provide accurate information about features, and that it can take five times as long to label instances as to label features. However, Raghavan, Madani, and Jones use the feature associations are used to perform feature selection, whereas in this paper we use such associations to train a probabilistic model, possibly without any labeled instances. In ongoing work we are extending the method used in this paper to perform active learning with features.

3 Generalized Expectation Criterion

A generalized expectation (GE) criterion objective function term assigns scores to values of a model expectation [McCallum *et al.*, 2007]. In many cases this score function is some measure of distance between a model expectation and a reference expectation. Specifically, given some distance function $\Delta(\cdot,\cdot)$, a reference expectation $\hat{f}$, an empirical distribution $\tilde{p}$, a function f, and a conditional model distribution p, the objective function is:

$$\Delta(\hat{f}, E_{\tilde{p}(X)}[E_{p(Y|X;\theta)}[f(X,Y)]]).$$

Here we use GE in conjunction with multinomial logistic regression models; $\Delta(\cdot,\cdot)$ is the KL divergence; D is unlabeled data U; and the expectations are predicted class distributions for unlabeled instances that contain input feature k, $\tilde{p}_k(y;\theta)$. We refer to this method as GE-FL. We define $\tilde{p}_k(y;\theta)$ as:

$$\tilde{p}_k(y;\theta) = \frac{1}{C_k}\sum_{\mathbf{x}\in U} p(y|\mathbf{x};\theta)x_k,$$

where $C_k = \sum_{\mathbf{x}\in U} x_k$. Notice that if input features are binary, $\tilde{p}_k(y;\theta) = \tilde{p}(y|x_k = 1;\theta)$. The estimation of reference distributions $\hat{p}_k(y)$ from feature-class associations is discussed in Section 4. A single GE-FL objective function term is then:

$$D(\hat{p}_k(y)||\tilde{p}_k(y;\theta)) = \sum_y \hat{p}_k(y)\log\frac{\hat{p}_k(y)}{\tilde{p}_k(y;\theta)}\,.$$

The combined GE-FL objective function is composed of multiple GE terms and a Gaussian prior over parameters.

$$\mathcal{O} = -\sum_{k\in K} D(\hat{p}_k(y))||\tilde{p}_k(y;\theta)) - \frac{\sum_j \theta_j^2}{2\sigma^2},$$

where K is the set of all features with at least one association. We use gradient methods to estimate parameters [Malouf, 2002]. The gradient of a GE-FL objective term with respect to the model parameter for feature j and label y', $\theta_{y'j}$, is:

$$\begin{aligned}
&\frac{\partial}{\partial\theta_{y'j}} D(\hat{p}_k(y)||\tilde{p}_k(y;\theta)) \\
&= -\frac{\partial}{\partial\theta_{y'j}}\sum_y \hat{p}_k(y)\log\tilde{p}_k(y;\theta) \\
&= -\frac{1}{C_k}\sum_y \frac{\hat{p}_k(y)}{\tilde{p}_k(y;\theta)}\sum_{\mathbf{x}\in U} x_k\frac{\partial}{\partial\theta_{y'j}}p(y|x;\theta) \\
&= -\frac{1}{C_k}\sum_y \frac{\hat{p}_k(y)}{\tilde{p}_k(y;\theta)}\sum_{\mathbf{x}\in U} x_k\Big(I(y=y')p(y|x;\theta)x_j - p(y|x;\theta)p(y'|x;\theta)x_j\Big) \\
&= -\frac{1}{C_k}\sum_y \frac{\hat{p}_k(y)}{\tilde{p}_k(y;\theta)}\sum_{\mathbf{x}\in U} p(y|x;\theta)x_k\Big(I(y=y')x_j - p(y'|x;\theta)x_j\Big)
\end{aligned}$$

Since there are more parameters in the model than corresponding GE terms in the objective function, the problem is under-constrained, and we expect there will be many optimal parameter settings. The Gaussian

prior addresses this problem by preferring parameter settings with many small values over settings with a few large values. This encourages the model to have non-zero values for parameters that do not have any labels, but co-occur with a labeled feature often. Above, we can see that the degree to which the gradient of a parameter for an unlabeled feature j and label y' is affected by a GE-FL term for labeled feature k depends on how often j and k co-occur in an instance.

Notice that if the distributions $\hat{p}_k(y)$ and $\tilde{p}_k(y;\theta)$ match exactly, then the gradient is zero:

$$\begin{aligned}
&\frac{\partial}{\partial\theta_{y'j}} D(\hat{p}_k(y)||\tilde{p}_k(y;\theta)) \\
&= -\frac{1}{C_k}\sum_y \frac{\hat{p}_k(y)}{\tilde{p}_k(y;\theta)} \sum_{x\in U} p(y|x;\theta)x_k\Big(I(y=y')x_j - p(y'|x;\theta)x_j\Big) \\
&= -\frac{1}{C_k}\sum_{x\in U}\sum_y p(y|x;\theta)x_k\Big(I(y=y')x_j - p(y'|x;\theta)x_j\Big) \\
&= -\frac{1}{C_k}\Big(\sum_{x\in U} x_k x_j \sum_y p(y|x;\theta)I(y=y') - \sum_{x\in U} p(y'|x;\theta)x_k x_j \sum_y p(y|x;\theta)\Big) \\
&= -\frac{1}{C_k}\Big(\sum_{x\in U} p(y'|x;\theta)x_k x_j - \sum_{x\in U} p(y'|x;\theta)x_k x_j\Big) = 0
\end{aligned}$$

3.1 Connections to a Dirichlet Prior

Notice that when $\Delta(p_1, p_2)$ is the KL-divergence, the GE criterion objective function has the form of a Dirichlet prior on the model's predicted distribution on unlabeled data:

$$L(\theta;U) = \frac{1}{\beta(\hat{p}_k(y))}\prod_y \tilde{p}_k(y;\theta)^{\hat{p}_k(y)}$$

$$\log L(\theta;U) = -\beta(\hat{p}_k(y)) + \sum_y \hat{p}_k(y)\log\tilde{p}_k(y;\theta)$$

Specifically, the partial derivative $\frac{\partial}{\partial\theta_{y'j}}$ of $\log L(\theta;U)$ is the same as the partial derivative of the KL-divergence $D(\hat{p}_k(y)||\tilde{p}_k(y;\theta))$. We emphasize that this Dirichlet prior is over model expectations, rather than model parameters. In future work, we plan to explore this interpretation more thorougly.

4 Estimating Reference Distributions

The quality of the reference distributions $\hat{p}_k(y)$ is clearly important to the success of the method. We estimate reference distributions with the following process:

1. Generate a list of candidate features using the unlabeled data.
2. Allow the labeler to choose associations for each feature, or discard the feature.
3. Estimate reference distributions using all of the associations.

We prefer the annotator provides associations because it may not be reasonable to assume the annotator can provide accurate estimates of the reference distributions directly. Below, we discuss methods for each part of this process.

4.1 Candidate Feature Selection

Ideally, a selected feature should be both highly predictive of some class, and occur often enough to have a large impact. In practice we will not be able to determine whether a feature is predictive if we have no labeled instances. However, in order to obtain an upper bound on feature selection methods, we assume there exists an oracle that can reveal the label of each unlabeled instance. A simple metric to determine the predictive power of a feature given labeled data is the entropy of the conditional distribution $\tilde{p}_k(y) = \frac{1}{C_k} \sum_{x \in U} \tilde{p}(y|x)x_k$. We select candidate features that occur in more than m instances and have low $H(\tilde{p}_k(y))$. We refer to this method as *oracle-features*.

Another potential feature selection method would select features randomly only according to their frequency. The problem with this method is that it tends to select common, non-predictive features, such as stopwords in text classification. Instead we use unsupervised feature clustering and select the most prominent features in each cluster. In this paper we cluster unlabeled data with Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003], a widely used topic model. For each LDA topic t_i, we sort features x_k by $p(x_k|t_i)$ and choose the top f features. There is no guarantee that the candidate features selected by this heuristic are relevant to the learning task of interest. However, in practice this performs much better than selecting candidate features by frequency. We refer to this method as *lda-features*.

4.2 Obtaining Feature-Class Associations

Since human provided associations are difficult to obtain, for some experiments we assume an oracle provides associations. Given a feature, the oracle correctly returns the associated classes, or, if there are no associated classes, discards the feature. We say that a feature is associated with a class if $\tilde{p}_k(y)$ is sufficiently greater than uniform. The oracle can reveal the labels for unlabeled examples, so $\tilde{p}_k(y)$ can be computed exactly. We stress however that the oracle only returns the association, not $\tilde{p}_k(y)$. We refer to this method as *oracle-associations*.

The second method for obtaining associations is to ask real annotators. We explore this approach in Section 5.4. We refer to this method as *human-associations*.

4.3 Reference Distribution Estimation

Assuming we have an oracle that reveals the labels of the unlabeled instances, we can estimate the reference distribution directly using the labeled data, without using the associations. This method, which we refer to as *oracle-distributions*, provides an upper bound for the other reference distribution estimation methods.

As proposed by Schapire, et al. [2002], we also apply a simple heuristic in which a majority of the probability mass is distributed uniformly among the associated classes(s), and the remaining probability mass is distributed uniformly among the other non-associated classes. Define q_{maj} as the probability for the associated classes. Then, if there are n associated classes out of L total classes, each associated class has probability $\hat{p}_k(y) = q_{maj}/n$ and each non-associated class has probability $\hat{p}_k(y) = (1 - q_{maj})/(L - n)$. For the experiments in this paper, we use $q_{maj} = 0.9$. We refer to this method as *Schapire-distributions*.

Alternatively, we use the feature-class associations to vote on labels for the unlabeled data. For each feature x_k in an instance x, it contributes a vote for each of its associated classes. We then normalize the vote totals to get a distribution over classes for each instance, and use this distribution to soft-assign labels. At this point, we can estimate the reference distributions directly using the soft labeled data. We refer to this method as *association-voted-distributions*.

5 Experimental Results

We evaluate GE-FL on text classification tasks. For all tasks, instances correspond to documents and features are word counts. We use three standard data sets. In the **20 newsgroups**[1] dataset, the task is to classify

[1] Available for download at: `http://www.cs.umass.edu/~mccallum/code-data.html`

dataset	baseline 1	baseline 2	baseline 3	xr
med-space (100)	0.87	0.82	0.84	**0.97**
ibm-mac (100)	**0.86**	0.84	0.82	**0.86**
financial-healthcare (100)	0.87	0.83	0.86	**0.91**
baseball-hockey (100)	0.90	0.82	0.89	**0.95**
webkb (200)	0.70	0.76	0.73	**0.78**
20 newsgroups (1000)	0.54	0.50	**0.57**	0.54

Table 1: Comparing methods that leverage feature-class associations.

messages according to the newsgroup to which they were posted. We use both the entire dataset (20,000 instances) and two-class subsets (2,000 instances) in the experiments below. We also use the **webkb**[2] dataset (4,199 instances), in which the task is to classify webpages as *student*, *course*, *faculty*, or *project*. Finally, we use the **financial vs. healthcare** subset (1,394 instances) of the **industry sector**[3] dataset, in which the task is to classify webpages according to a hierarchy of industrial sectors. For all text classification experiments, we simulate unlabeled data by hiding labels of all instances.

5.1 Comparison with Baseline Methods

We compare GE-FL with several baseline methods, described below. For these experiments, we use *oracle-associations*.

- (1) Use the provided feature-class associations to choose the class with the largest number of associated features in each instance.
- (2) Use GE to match reference distributions estimated using *association-voted-distributions*, but disallow parameter changes for features without associations.
- (3) Compute a soft labeling of the unlabeled data with *association-voted-distributions*, and then train a supervised multinomial logistic regression model directly on this data.

We run experiments comparing the above baselines with GE-FL and provide the results in Table 1. Datasets **med-space**, **ibm-mac**, and **baseball-hockey** are subsets of the **20 newsgroups** dataset, whereas **healthcare-financial** is a subset of the **industry sector** dataset. The parenthesized number indicates the total number of features with associations. Candidate features are selected using *oracle-features*. For GE-FL, reference distributions are estimated using *association-voted-distributions*. GE-FL outperforms other methods (by as much as 10%) on all data sets except **20 newsgroups**.

Each of these baselines demonstrates an important point about GE-FL. Baseline 1 uses the prior knowledge only, whereas GE-FL matches prior knowledge with model expectations on unlabeled data, and in the process learns about co-occurring features without associations. Baseline 2 demonstrates the importance of incorporating these co-occurring features without associations. Finally, Baseline 3 uses prior knowledge to infer reference distributions on all features, whereas GE-FL only specifies constraints on features with known associations.

5.2 Comparison with Schapire, Rochery, and Gupta [2002]

Next, we compare against the method proposed by Schapire, et al. [2002], on the (full) **20 newsgroups** data set. Schapire, et al. provide several associated features for each class, 138 in total. Some features appear as indicators of multiple classes. The distributions over classes for each feature used in the boosting algorithm are estimated using the *Schapire-distributions* heuristic. We note that Schapire et al. use n-gram features, whereas we use only unigram features.

[2]Available for download at: `http://www.cs.cmu.edu/~webkb`
[3]Available for download at: `http://www.cs.umass.edu/~mccallum/code-data.html`

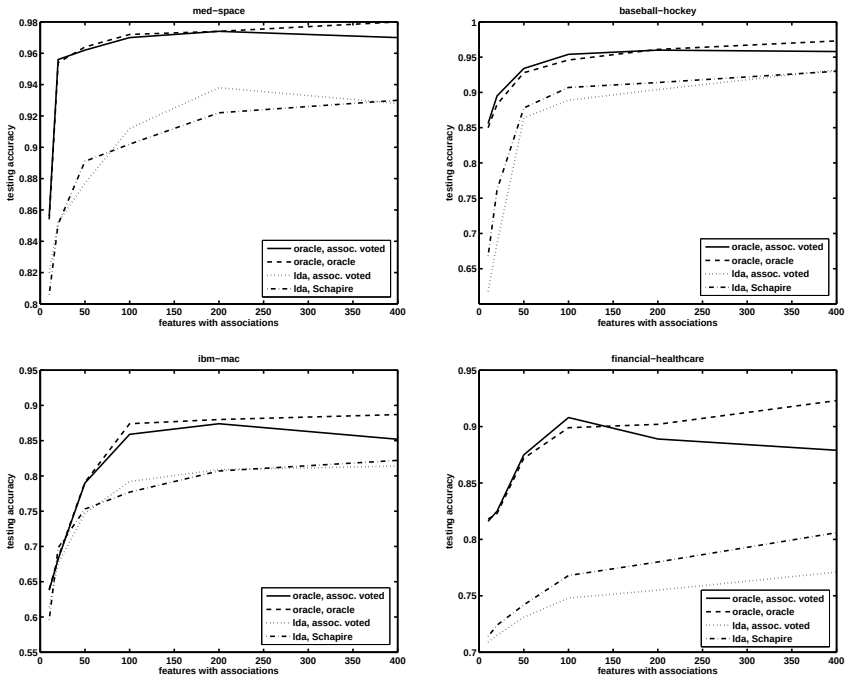

Figure 1: Accuracy vs. the number of feature associations using combinations of methods for selecting candidate features and estimating reference distributions. Specifically, the first item in the legend indicates the candidate feature selection method, and the second indicates the method for estimating reference distributions.

To ensure a fair comparison, we use the same list of associations and the same technique to estimate the distributions. Importantly, we treat the association list as the output of feature labeling, so we do not discard any features. Using the associations only, the GE-FL method gives approximately a 10% absolute error reduction from 64% error ([Schapire *et al.*, 2002]) to 54% error. This improvement comes from matching class distributions on unlabeled data, which implicitly learns soft associations for features without associations. It requires an additional 400 labeled documents for the method of Schapire et al. to match the accuracy of GE-FL, which uses no labeled documents.

5.3 Feature Selection and Prior Estimation

Here we compare the heuristics for selecting candidate features and for estimating priors described in Section 4. For these experiments, we use *oracle-associations*. Figure 1 shows that although the oracle methods clearly perform better, we can still train accurate models using the simple heuristic methods. Figure 1 also shows that we need a relatively small number of associations to obtain near best performance.

5.4 User Experiments

Finally, we perform annotation experiments in which we time three users as they label 100 documents and 100 features for binary classification tasks. The candidate features are selected using *lda-features*. The features are presented one at a time, and the user can choose an associated class for the feature or choose to discard the feature. After the users finish selecting features, they label documents, again being able to choose the label of the document or ignore the document if it appears ambiguous. We prefer this ordering of labeling features followed by documents in order to give maximum benefit to the traditional document

medical vs. space	nasa (1), blood (0), planet (1), universe (1), medicine (0), health (0), ...
ibm vs. mac	apple (1), hp (0), dos (0), mac (1), ibm (0)
baseball vs. hockey	nhl (1), flyers (1), ball (0), braves (0), goal (1), runs (0), hitter (0), ...

Table 2: Features for which all three labelers gave the same (correct) association.

	doc labeling	feat selection	feat labeling
User 1 ibm-mac	0.90	0.80	1.00
User 1 med-space	0.95	1.00	1.00
User 1 baseball-hockey	0.98	1.00	0.91
User 2 ibm-mac	0.92	0.88	0.75
User 2 med-space	0.98	0.93	0.86
User 2 baseball-hockey	0.96	0.88	0.79
User 3 ibm-mac	0.91	0.86	1.00
User 3 med-space	0.99	1.00	0.97
User 3 baseball-hockey	0.96	1.00	1.00
Overall mean	0.95	0.93	0.92

Table 3: User labeling accuracy with respect to oracle for document labeling, feature selection, and feature labeling.

labeling method. We choose documents to present to the user with uncertainty sampling: after each instance is labeled, the instance with the most uncertain classification under the current model is selected next for labeling. In our experiments this is superior to randomly choosing documents.

In Table 2, we show features for which all annotators agreed. For example, all annotators chose to label word feature "nasa" as indicative of the *space* class for the **med-space** dataset. Table 3 shows the labeling accuracies for different annotators. Feature selection accuracy refers to the proportion of features that are not discarded that are sufficiently predictive for one of the classes ($\exists y$ such that $\tilde{p}_k(y) > 0.6$). Feature labeling accuracy refers to the proportion of features whose associated class labeling is correct (if y is given, $\tilde{p}_k(y) > 0.5$). All three annotators were able to label the documents well. User 3 had trouble selecting and labeling features. We suspect that this indicates insufficient familiarity with the learning tasks. In fact, several other users, whose results are not reported here, were not able to complete the labeling (of either features or documents) as a result of lack of knowledge about the tasks. This suggests that future experiments should involve an opportunity to look through the data before annotation. However, assuming that the annotators are familiar with the task they are trying to solve does not seem unreasonable.

Figure 2 shows the accuracy of two trained systems over time. The first uses the labeled features and unlabeled instances with GE-FL. Reference distributions are estimated using *Schapire-distributions* with $q_{maj} = 0.9$. The second uses entropy regularization (ER) [Grandvalet and Bengio, 2004] with the labeled and unlabeled instances. It is clear from the graphs that annotating features yields significant accuracy improvements for the same amount of time. On average across all experiments, labeling features is 3.7 times faster than labeling documents, and the models trained with GE-FL have 1.0% higher final accuracy. Note that the point at which the GE-FL curve changes from a dotted line into dots indicates the point at which the user had processed all 100 features.

When the annotator is accurate, the results with feature labeling can be quite striking. For example, consider the results of User 1 for the **ibm vs. mac** classification task. The accuracy of the GE-FL system after 30 seconds of feature labeling is better than the accuracy of the ER system after 12 minutes of document labeling, a 24x speed-up. As another example, User 3 achieves accuracy of 90% on the **baseball vs. hockey** task after 90 seconds with the GE-FL system, at which point the ER system accuracy is around 50%.

Notice that the ER system gives erratic performance, with large accuracy jumps in consecutive 30 second intervals. This reinforces our earlier assertions about the brittleness of current semi-supervised methods.

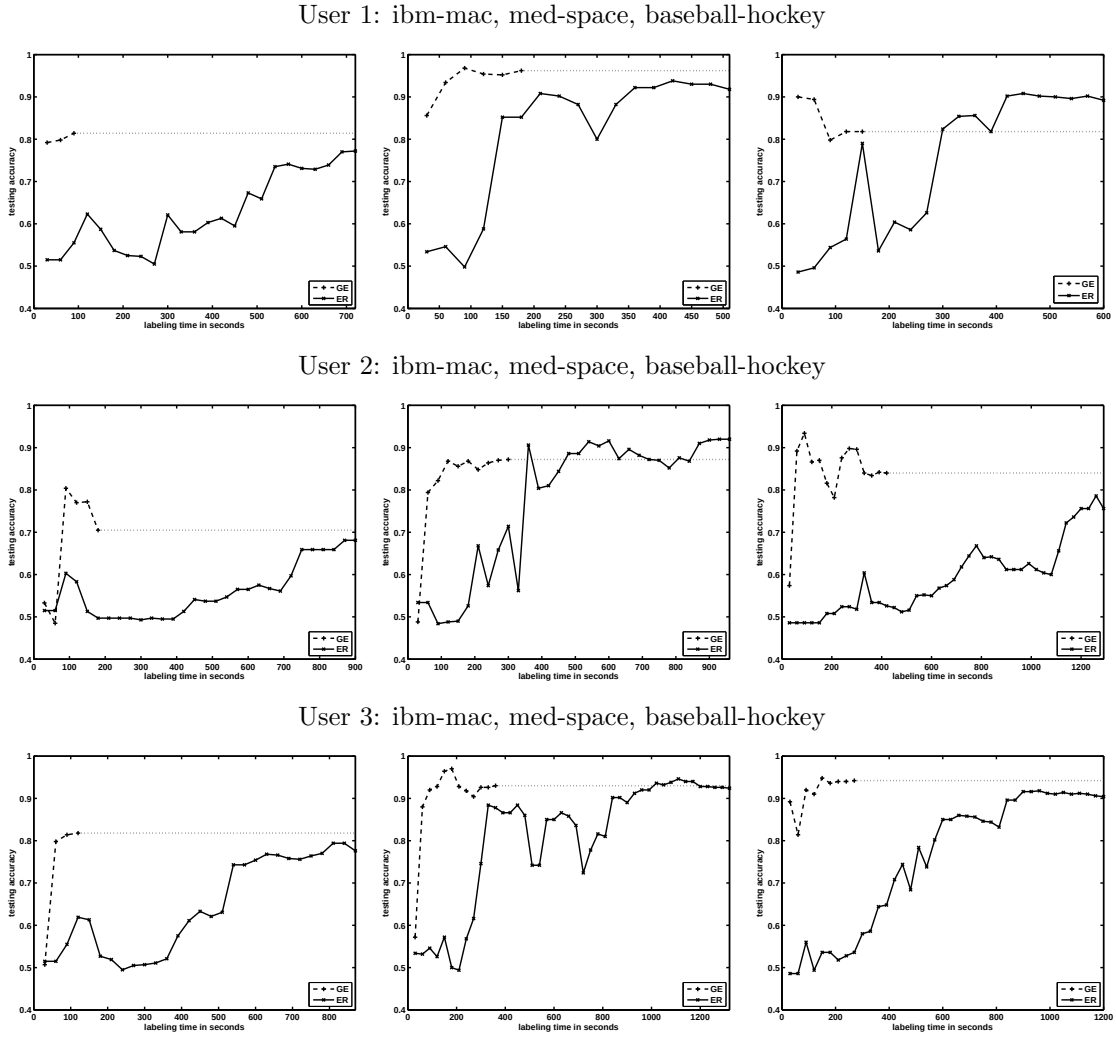

Figure 2: Accuracy vs. time for the GE-FL and ER systems. We find that in most cases, GE-FL gives better accuracy given the same amount of annotation time.

6 Conclusion and Future Work

We have used generalized expectation criteria to train multinomial logistic regression models using only feature-class associations and unlabeled data. The ability to leverage feature-class associations using GE provides an alternate modality of supervision in which instead of labeling instances as to their class, an annotator labels features with their associated classes. In our experiments, labeling features is faster than labeling instances, and training a model with GE-FL is more accurate than training a model with labeled and unlabeled instances using both traditional active and semi-supervised learning.

In ongoing research, we are working to improve reference distribution estimation, applying GE to models for structured output spaces, and applying GE to active learning and domain adaptation.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by DoD contract #HM1582-06-1-2013, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[Bengio *et al.*, 2006] Y. Bengio, O. Dellalleau, and N. Le Roux. Label propagation and quadratic criterion. In O. Chapelle, B. Schlkopf, and A. Zien, editors, *Semi-Supervised Learning*. MIT Press, 2006.

[Berger *et al.*, 1996] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, 1996.

[Blei *et al.*, 2003] D. M. Blei, A. Y. Ing, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[Chang *et al.*, 2007] M. Chang, L. Ratinov, and D. Roth. Guiding semi-supervision with constraint-driven learning. In *Proc. of the Annual Meeting of the ACL*, pages 280–287. Association for Computational Linguistics, 2007.

[Cohn *et al.*, 1994] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[Cozman and Cohen, 2006] F. Cozman and I. Cohen. Risks of Semi-Supervised Learning. In O. Chapelle, A. Zien, and B. Scholkopf, editors, *Semi-Supervised Learning*. MIT Press, 2006.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Stat. Soc.*, 39:1–38, 1977.

[Freund *et al.*, 1997] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[Graca *et al.*, 2008] J. Graca, K. Ganchev, and B. Taskar. Expectation maximization and posterior constraints. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, 2008.

[Grandvalet and Bengio, 2004] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, 2004.

[Haghighi and Klein, 2006] A. Haghighi and D. Klein. Prototype-driver learning for sequence models. In *NAACL*, 2006.

[Jin and Liu, 2005] R. Jin and Y. Liu. A framework for incorporating class priors into discriminative classification. In *PAKDD*, 2005.

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

[Lewis and Catlett, 1994] D.D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *ICML*, 1994.

[Malouf, 2002] R. Malouf. A comparison of algorithms for maximum enotrpy parameter estimation. In *COLING*, 2002.

[Mann and McCallum, 2007] G. Mann and A. McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML*, 2007.

[McCallum *et al.*, 2007] A. McCallum, G. Mann, and G. Druck. Generalized expectation criteria. Technical Report 2007-62, University of Massachusetts, Amherst, 2007.

[Raghavan *et al.*, 2006] H. Raghavan, O. Madani, and R. Jones. Active learning with feedback on both features and instances. *JMLR*, 2006.

[Schapire *et al.*, 2002] R. Schapire, M. Rochery, M. Rahim, and N. Gupta. Incorporating prior knowledge into boosting. In *ICML*, 2002.

[Wu and Srihari, 2004] X. Wu and R. K. Srihari. Incorporating prior knowledge with weighted margin support vector machines. In *ACM SIGKDD*, 2004.

[Zhu and Ghahramani, 2002] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, CMU, 2002.

[Zhu *et al.*, 2003] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic mixtures. In *ICML*, 2003.

[Zhu, 2005] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.