Learning with Scope, with Application to Information Extraction and Classification

David M. Blei
Computer Science Division
U.C. Berkeley
blei@cs.berkeley.edu

J. Andrew Bagnell
Robotics Institute
Carnegie Mellon University
dbagnell@ieee.org

Andrew K. McCallum
WhizBang Labs - Research
& Carnegie Mellon University
mccallum@whizbang.com

Abstract

In probabilistic approaches to text classification and information extraction, one typically builds a statistical model of words under the assumption that future data will exhibit the same regularities as the training data. In many text data sets however, there are additional, scope-limited features whose predictive power is applicable only to a certain subset of the data. For example, in information extraction from web pages, word formatting may be indicative of extraction category in different ways on different web pages. The central difficulty with using such features is capturing and exploiting the new regularities encountered in previously unseen data. In this work, we propose a hierarchical probabilistic model that uses both local/scope-limited features (e.g., formatting) and global features (e.g., word content). The local regularities are represented as an unobserved random parameter for each local data set, and these regularities are captured in the inference process. This process is akin to automatically retuning our classifier to the local regularities on each newly encountered web page. Exact inference is intractable, and we present approximations via point estimates and variational methods. Empirical results on large collections of web data show this method significantly improving performance over traditional models of global features alone.

1 Introduction

A fascinating aspect of classification and extraction from the web is the richness of the information in its formatting, layout, directory structures, and linkage. Within a particular web site, these structural regularities are often powerful features for common classification and extraction tasks. For example, if one wished to extract the titles of all the books on Amazon.com, one can rely on the fact that on the main page for each book, the book title always appears in the same location and in the same font. The difficulty with using this information is that each web site has its own, different structural regularities, and thus one cannot successfully apply models built for one site to extraction or classification from another. In response, many people have created tools to facilitate hand-tuning of site-specific extractors (Cohen and Jensen, 2001).

The chief obstacle in modeling these types of site-specific regularities without hand-intervention is that most statistical models assume data to be independent and identically distributed (IID). However, as in the above example, it is often the case that certain disjoint subsets of the data share identifiable and useful regularities that do not occur throughout its entirety. Some other examples of subsets that may have local regularity include patients from a particular hospital, voice sounds from a particular speaker, or vibration data from a particular airplane.

Leveraging these local regularities can significantly improve the performance of a learner because local features are often both simple and highly indicative. The central difficulty is that our trained algorithm will be applied to new subsets of the data with local regularities that were not encountered in training. The only knowledge directly applicable to the new data is from the traditionally-used, global regularities i.e., those features that are independent and identically distributed across all data.

This paper presents *scoped learning*, a probabilistic framework that takes advantage of local regularities in previously unseen data. We introduce a new graphical model within this framework for which some parameters are globally learned across all the training data and other parameters (for the particular local subsets

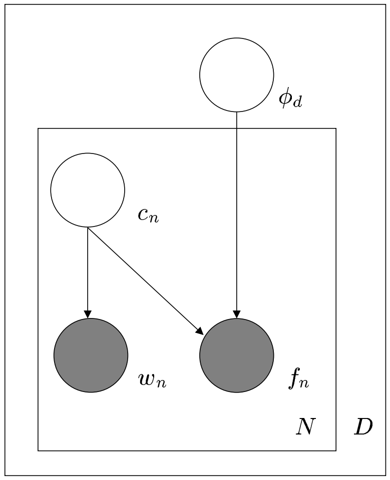

Figure 1: The graphical model representation of the scoped-learning model in a text domain. The square plates represent repetitions of documents (D) and the words in them (N). The (global) word content w_n is generated independently of any parameters local to each document. The local formatting feature f_n is generated dependent on the document-specific parameters ϕ. These parameters, in turn, are generated once per document.

of the data) are represented as unobserved random variables. Inference on this model naturally leads to a process by which we learn the local regularities of new subsets of data based on the information inferred from the global features. In this work we describe both generative and discriminative approaches to training and using the model.

2 A generative model of learning with scope

In this section, we describe a generative model of multi-dimensional data where the components of the data exhibit two kinds of scope. Some of the components, called *global* features, are governed by parameters which are shared for all datapoints. Other components, called *local* features, are governed by parameters which are shared for a particular subset of the datapoints.

A subset for which the local features exhibits some kind of regularity is called a *locale*. We assume that the future data on which we use our model is naturally organized in similar such locales. Furthermore, the regularities with respect to the local features in the unseen data may be very different from the local regularities in the data used for parameter estimation.

For simplicity of explanation, we will describe our model with reference to information extraction (IE) on web pages where the data naturally exhibits the kind of structure and regularities which our model can exploit. Note however that this model can easily be applied to any kind of data which exhibits scope in the way we described above.

We cast IE as a classification problem in which the goal is to label certain individual words on a web page with different extraction categories such as "Job Title" or "Book Review". The datapoints, specifically word content and its corresponding formatting, can be easily divided into subsets that form different locales according to page boundaries. For training, we are given data from multiple locales with an observed class label for each word-formatting pair. Future data are unlabeled web pages (i.e., new locales) for which we would like to label each word-formatting pair with the most likely extraction category.

The global features of web data for IE are the individual words on the page without reference to their formatting. This type of feature is a good predictor of extraction category in the same way for all documents. For example, the word "engineer" is probably part of a job title whether it appears on a page on www.whizbang.com or www.google.com.

The local features in this domain are the formatting of the individual words. These features, such as boldness or font size, can be a good predictor of category on a particular web page. However, the regularity which they capture does not necessarily hold on other pages in the data set. For example, all the job titles on www.whizbang.com may be green while all the job titles on www.google.com may be red.

A traditional probabilistic classification approach to this problem would treat all the word-formatting pairs as IID from some probabilistic model. With this approach however, the formatting features have very limited predictive power since in estimating the parameters, we cannot learn the specialized meaning of the local features in future domains. In contrast, the model developed here can successfully exploit the formatting features in yet unseen pages.

2.1 Generative process and parameterization

The key problem in constructing a generative model of local and global features is that the word content (global features) and formatting (local features) are not IID since their joint distribution depends on the particular document to which each individual pair belongs. However, we can model *documents* as IID by introducing an unobserved random parameter that governs the local features for a particular document. In particular let an N word document contain corresponding word content $\mathbf{w} = \{w_1, \ldots, w_N\}$, formatting $\mathbf{f} = \{f_1, \ldots, f_N\}$, and class labels $\mathbf{c} = \{c_1, \ldots, c_N\}$. We assume that the document is generated by the following process:

1. For each of the K extraction categories:
   (a) Generate the formatting feature parameters ϕ_i from $p(\phi_i)$.
2. For each of the N words in the document:
   (a) Generate the nth class label c_n from $p(c_n)$.
   (b) Generate the nth word from $p(w_n|c_n)$.
   (c) Generate the nth formatting feature from $p(f_n|c_n, \phi)$.

This models the following joint distribution over local parameters, class labels, words, and formatting features:

$$p(\phi, \mathbf{c}, \mathbf{w}, \mathbf{f}) = p(\phi) \prod_{n=1}^{N} p(c_n) p(w_n|c_n) p(f_n|c_n, \phi),$$

which is exhibited as a graphical model in Figure 1. Let K be the number of extraction categories, V be the size of the vocabulary, and F be the different possible values of the formatting feature. The parameters of this model are: K multinomials over V words for $p(w|c)$; a multinomial over K categories for $p(c)$; and K distributions over the random parameters ϕ_i.

Note that for each document and extraction category, ϕ_i is a random parameter to a multinomial over values of the formatting feature, i.e., $p(f_n|c_n, \phi) = \phi_{c_n, f_n}$. Each random parameter then is a point on the F-simplex which we can parameterize by a Dirichlet distribution (recall that the Dirichlet is the exponential family distribution on the simplex). Thus, we need $K \times F$ Dirichlet parameters for the full distribution on the document-specific parameters.

It is important to observe that there are no parameters which explicitly reference how the formatting features are generated for a particular document – only parameters on the random document-specific parameters which govern those features. This is a key element to the model since the formatting features will be generated by different parameters for different documents.

Finally, note that we assume that the local features and global features are independent of each other given the class label. This is often a reasonable assumption. In IE, the formatting of a word depends more on its role in the document (i.e., the extraction category) than what that word is in particular.

2.2 Training

We fit our model by maximum likelihood estimation on a set of documents for which all the class labels are observed. Given the label, the (global) words and (local) formatting features are independent. Thus the global parameters $p(w|c)$ and $p(c)$ can be estimated without knowledge of formatting or how the words are divided into documents. In other words, we can train the class-conditional word distributions and class prior as a traditional generative probabilistic classifier (e.g., naive Bayes).

The other global parameters are the Dirichlet parameters on $p(\phi_i)$. In learning these we can quantify general trends in the local contexts (e.g., "boldness is important" or "font size 12 is usually not indicative of class label"). In our experiments however, we fix the parameters to give a uniform distribution on the random local distributions. With these parameters fixed, we do not need to estimate any parameters with the formatting features of the training set. The formatting of future data, by the definition of the model, is generated by different parameters than those that generated the training data.

2.3 Inference

The power of our scoped-learning model lies in its inference routines. Given a new web page, we would like to classify its word-formatting pairs based on both which words they are and how they are formatted with respect to the rest of the document. In our model, inference routines on the class labels naturally lend themselves to a process by which we can leverage what we know globally in order to infer the document-specific local parameters. Then, we can refine our original globally-estimated labels using the local features.

The key difference between the inference routines of the scoped-learning model and a standard classifier is that we aim to simultaneously infer all the information labels of the document at once rather than treating each word-formatting pair separately. These labels are dependent on each other through the unobserved local parameter.

In particular, we would like to compute the posterior distribution on the set of information labels $\mathbf{c}$ for some document ($\mathbf{w}$,$\mathbf{f}$). In principle, this can be computed by marginalizing out the document-specific parameter ϕ and invoking Bayes rule:

$$p(\mathbf{c}|\mathbf{w}, \mathbf{f}) = \frac{\int \prod_{n=1}^{N} p(w_n|c_n) p(f_n|c_n, \phi) p(c_n) p(\phi) d\phi}{\int \prod_{n=1}^{N} \sum_{c_n} p(w_n|c_n) p(f_n|c_n, \phi) p(c_n) p(\phi) d\phi}. \quad (1)$$

Exact computation of the integral is computationally infeasible because of the denominator, an integral over the simplex of a product of sums. We present two approaches to approximate inference below. Both approaches operate on the model induced by the single document in question. One approach is maximum likelihood estimation of ϕ on that model. The other, more Bayesian approach, uses variational methods to approximate the infeasible integral in Equation 1.

2.3.1 MAP estimates of the local parameter

If we approximate ϕ with a point estimate $\hat{\phi}$, the posterior on $\mathbf{c}$ is tractable. To see this, observe that the integral in Equation 1 disappears and replace each instance of ϕ with $\hat{\phi}$. This yields the labels for each word-formatting pair independent of each other given $\hat{\phi}$. We can therefore label each pair with:

$$\hat{c}_n = \arg\max_{c_n} p(w_n|c_n)p(f_n|c_n, \hat{\phi})p(c_n).$$

A natural point estimate for the local parameter is the posterior mode or MAP estimate, $\hat{\phi} = \arg\max_\phi p(\phi|\mathbf{f}, \mathbf{w})$. This is simply a maximum likelihood estimate of the local parameters where we maximize the likelihood of the single document in question and hold the global parameters fixed.

Since $\mathbf{c}$ is a set of unobserved random variables, we can use the Expectation-Maximization (EM) algorithm to maximize the expected log likelihood of the formatting features of the document:

$$\mathcal{L}_d(\phi) = \sum_{n=1}^{N} \sum_{c_n=1}^{K} p(c_n|w_n, f_n; \phi) \log p(f_n|c_n; \phi). \quad (2)$$

The EM algorithm is coordinate ascent in this function. The E-Step computes the posterior distribution over extraction categories:

$$p^{(t+1)}(c_n|w_n, f_n; \phi) \propto p^{(t)}(f_n|c_n; \phi)p(w_n|c_n)p(c_n).$$

The M-Step finds the new estimate of $\hat{\phi} = p(f|c)$:

$$p^{(t+1)}(f|c; \phi) \propto \sum_{\{c_n : c_n = c\}} p^{(t)}(c_n|\ell_n, w_n),$$

where the notation under the sum indicates all training instances such that $c_n = c$. Alternating between these two steps guarantees that we take positive steps in Equation 2. We can then use the resulting $\hat{\phi}$ to find the best value of c_n for each word-formatting pair.

2.3.2 Variational approximation

As an alternative to using a point estimate, we can approximate the true posterior distribution on the class labels via variational methods. With this approach, we can better approximate the actual integrals in Equation 1.

In particular, we define a factorized variational distribution for each new document on the unobserved random variables $\mathbf{c}$ and ϕ:

$$q(\mathbf{c}, \phi) = \prod_{k=1}^{K} q(\phi_k; \gamma_k) \prod_{n=1}^{N} q(c_n; \mu_n),$$

where γ_k are a set of variational Dirichlet parameters for each extraction category and μ_n are a set of variational multinomial parameters over categories for each word-formatting pair. These are free parameters which we optimize to find the variational distribution that is as close as possible, in KL divergence, to the true posterior:

$$(\hat{\gamma}, \hat{\mu}) = \arg\max_{(\gamma,\mu)} \mathrm{KL}(q(\mathbf{c}, \phi; \gamma, \mu) \| p(\mathbf{c}, \phi|\mathbf{w}, \mathbf{f})).$$

This is a very similar framework to Blei et al. (2002) and yields similar updates of the variational distribution.

Let γ_{ij} be the Dirichlet parameter for the jth value of f and the ith information category. The variational Dirichlet parameters are maximized by:

$$\gamma_{ij} = 1 + \sum_{\{f_n : f_n = j\}} q(c_n = i; \mu_n).$$

Note that this is similar to a posterior Dirichlet distribution with a uniform prior and expected counts under the variational distribution.

The variational multinomial parameters are updated by:

$$q(c_n; \mu_n) \propto p(c_n)p(w_n|c_n) \exp(\mathrm{E}[\log p(f_n; \phi); \gamma_c]).$$

Note that this equation is similar to the E-step in Section 2.3.1 with $p(f_n|\hat{\phi}, c_n)$ replaced by the exponential of the expected value of its log under the posterior (variational) Dirichlet. We can easily compute this expectation:

$$\mathrm{E}[\log p(f_n; \phi); \gamma_c] = \Psi(\gamma_{cf_n}) - \Psi(\textstyle\sum_f \gamma_{cf}),$$

where Ψ is the digamma function. Iterating between the updates of γ and μ will converge on the variational parameters which are closest to the true posterior. It is important to notice that while the variational approximation may mitigate the problem of the EM algorithm overfitting inside a locale, it maintains the same computational complexity and implementation simplicity.

We can view this attempt to approximate the integral over ϕ from a Bayesian point of view. In particular, a point estimate of the local parameters is only justified in large data limits where $p(\phi|\mathbf{w},\mathbf{f})$ will be very peaked at a particular value of ϕ. Since we are specifically modeling small subsets of word-formatting pairs, this large data assumption will generally be invalid and thus we can expect to gain from this attempt at proper integration. With this perspective, we can see that the update equations for the variational distribution follow the variational Bayes recipe given in Attias (2000).

3 A discriminative model of learning with scope

We have assumed in Section 2 that we are training a generative model of the local (formatting) features. In some cases a generative model places a strong and unwarranted independence assumption on the local features and it may be desirable to learn a local discriminative model that captures $p(\mathbf{c}|\mathbf{f})$. Graphically, we model this by reversing the arc from f to c and pointing the random parameter ϕ to the category node c in Figure 1.

As before, exact inference is intractable. We can find a point estimate for the optimal conditional parameters by maximizing the conditional log likelihood $p(\mathbf{w}|\mathbf{f};\phi)$:

$$\log p(\mathbf{w}|\mathbf{f};\phi) = \sum_{n=1}^{N} \log \sum_{c_n} p(c_n|f_n;\phi)p(w_n|c_n).$$

Note that we are implicitly assuming a generative global classifier. To recover a completely discriminative solution, we use Bayes' rule to rewrite $p(w_n|c_n)$ as $p(c_n|w_n)p(w_n;\phi)/p(c_n;\phi)$. This rewriting exposes a dependence of the likelihood on the marginal probability of the global features which we wish to avoid modeling. However, by enforcing the constraint that $p(\mathbf{c})$ is constant with respect to ϕ, $p(\mathbf{w})$ also is independent of ϕ and thus factorizes out of the likelihood. We then optimize the local parameters with respect to the following complete data likelihood:

$$\mathcal{L}_{\text{cond}}(\phi) = \sum_{n=1}^{N} \sum_{c_n} p(c_n|w_n,f_n) \log \frac{p(c_n|f_n;\phi)p(c_n|w_n)}{p(c_n)}.$$

The EM algorithm derivation follows as in the MAP approximation. The E-step, again, computes the posterior distribution:

$$p^{(t+1)}(c_n|w_n,f_n) \propto \frac{p^{(t)}(c_n|f_n)p(c_n|w_n)}{p(c_n)}.$$

The M-step corresponds to maximizing the weighted log likelihood:

$$J^{(t+1)} = \sum_{n=1}^{N} \sum_{c_n} p^{(t)}(c_n|w_n,f_n) \log p(c_n|f_n;\phi),$$

which is trivial to implement for most discriminative probabilistic classifiers (e.g., maximum entropy).

The conditional likelihood objective function $p(\mathbf{w}|\mathbf{f})$ has interesting information-theoretic content as well. We note that the criteria is equivalent to maximizing an empirical mutual information $\mathcal{I}(\mathbf{w},\mathbf{f};\phi)$. [1] We can thus consider the class variables and the local parameters to be a "channel" where we try to recover as much information as possible about the global features from the local features.

4 Example and empirical results

4.1 Data

We test the scoped-learning model on two real-world data sets obtained from the web. The first is a data set of 1000 HTML documents that are automatically divided into sets of words with similar layout characteristics. Each group of words (called a "text node") is hand-labeled as to whether it contained a job title. We split the data 50/50 into a testing set and training set, keeping text nodes of a single document and documents of a single site together. Given an unlabeled document, the goal of our model is to correctly label the text nodes. As we have described, the global features of this data are the word content and the local features are the formatting of the words. The formatting features are chosen using the formatting tags in the HTML source.

Our second data set consists of 42,548 web pages from 330 web sites where each page is hand-labeled as to whether it is a press release. Given a new site, our goal is to find the press-release pages within that site. The global features, again, are the words of the documents. The local features come from the URL. One expects that within a particular web site, the URL structure is consistent with the function of the document to which it refers. For example, all press release pages of a particular company may be in a subdirectory `company/news`, or may be in files named similarly to `pressrel20010402.html`. The particular local features which we used were all the 4-grams of characters that appear in the URL. As in the job title data set, we split the data 50/50 into training and testing sets.

We should note that the actual scoped-learning model which we used for both these data sets is slightly augmented from the one depicted in Figure 1. In particular, each class label is associated with multiple local and global features. Thus we have repetitions of the w

[1] By empirical mutual information, we mean $\sum \hat{p}(w,f) \log \frac{p(w,f;\phi)}{p(w)p(f)}$ where $\hat{p}(w,f)$ is the empirical distribution of the features.

and f nodes inside the repetition over class labels. The training and inference procedures however are practically the same so we chose to present them in the more simplified setting.

4.2 An example page

We present an example web page from the job title corpus to aid in understanding the scoped-learning model and the power we gain from inference on the local parameters. Figure 2 shows a webpage from the test set. Clearly, the job titles on this page have a consistent formatting.

A simple naive Bayes classifier mislabels 5 of the 8 jobs on this page due to the lack of training data. In particular, it misses "Family Services Director", "Massage Therapist", "Starbucks Server", "Teachers", and "Lifeguard". However, for all these job titles (with the exception of "Starbucks Server") the classifier is relatively unsure of the label and only leans slightly away from classifying them correctly. Furthermore, the global classifier is very sure of the jobs which it correctly labels, "Teacher", "Art Coordinator", and "Customer Service Rep".

When we apply the scoped-learning model to this document, we learn from the correctly labeled jobs that the probability of boldness, the Arial font, and orange is very high, conditioned on the JOB-TITLE class label. Posterior inference on the class labels then yields the correct labeling for all but "Starbucks Server"—giving a significant increase in classification accuracy. As we will see in the next section, this improvement is typical on this data set and we can dramatically improve overall classification performance.

4.3 Quantitative results

Since both of our data sets are binary classification problems, we can quantify performance with precision and recall. In particular, we can change the threshold of probability with which we classify an example as being the positive class. In doing so, we can trade off classifying more negative examples as positive and finding more true positive examples (i.e., low precision, high recall) and missing more true positives but reducing the number of false positives (i.e., high precision, low recall).

The graph in Figure 3 (left) illustrates the performance of the naive Bayes classifier and the two generative variants of inference on the scoped-learning model on the job title data set. The first variant is the MAP estimate of the local parameters learned via EM. We note that it consistently dominates the performance of the global classifer. The second variant is the variational inference algorithm. This algorithm also dominates the global classifier and generally dominates the MAP estimate as well. In particular, note that at high levels of recall, which is an important part of the graph from a practical perspective, both algorithms that exploit the local features show significant improvement over the global classifier. Another measure of performance is the traditional accuracy (percentage of correct labels given at 0.5 threshold); by this measure, the scoped-learning method reduces error by 66%.

The graph in Figure 3 (right) illustrates the performance of some of our algorithms on the press release data set. In this case, the global classifier and scoped-learning discriminative classifiers are both maximum entropy classifiers with a Gaussian prior on the parameters. Although it is relatively easy in this data set to obtain high precision at low and moderate recall, the more interesting task is to maintain such precision at higher recall. The scoped-learning models again outperform the global classifier at these levels of recall, reducing error by more than one third at 90% recall, and by more than half at 80% recall. Note that in this data set, the local features are highly dependent and thus the discriminative approach outperforms the MAP estimate.

5 Related work

The central focus of this work is to emphasize the utility of features which exhibit scope-limited regularity and methods of learning this regularity in previously unseen data.

One thread of related work is Blum and Mitchell's PAC-style co-training (1998) in which two independent views of the data—each sufficient to predict the class label without error—can influence each other in making a final prediction. Although co-training does not exploit scope-limited features, its conditional independence assumption is similiar to the one made here.

Work in probabilistic classification of relational data (Taskar et al., 2001) extends the notion of multiple views of a data set to multiple kinds of relationships and connections between the components of the data. The authors use probabilistic relational models (PRM's) to represent an entire data set as an interconnected graphical model. Inference on the unknown class variables then implicitly exploits the regularities learned about different kinds of features which are present in the data.

In a sense, the models presented here can be cast as a PRM where each locale would be represented as a separate group of nodes with a separate local parameter. This representation, and the maximum likelihood

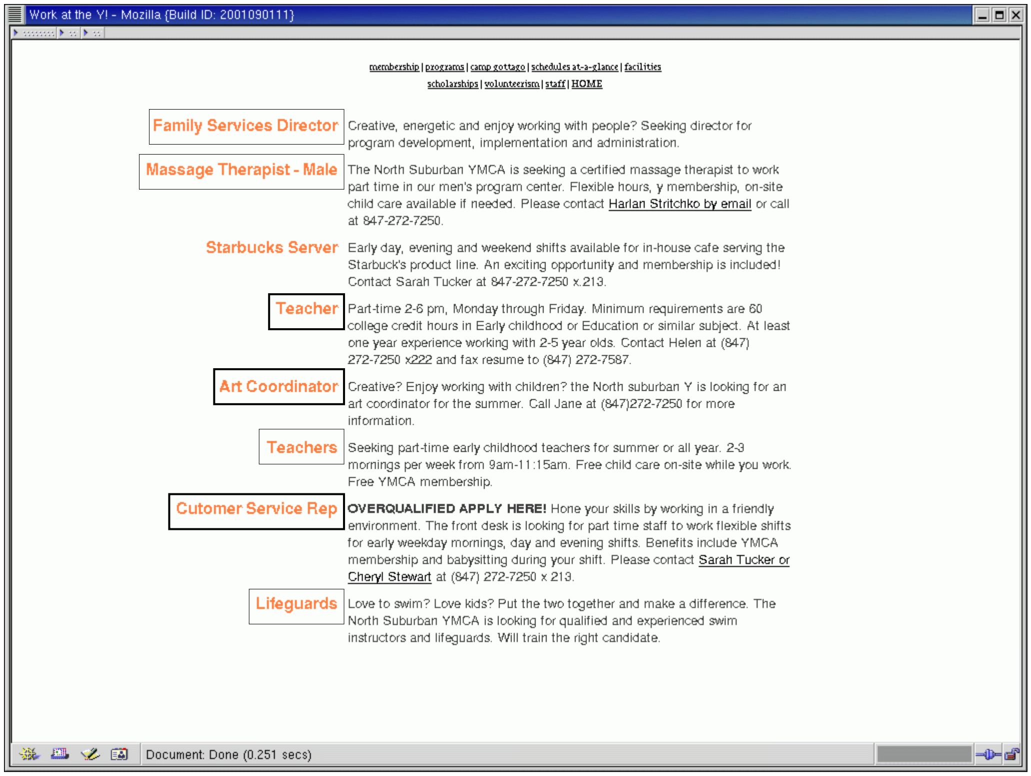

Figure 2: A sample webpage from the job title dataset. The global classifier (depicted by the bold rectangles) correctly labeled 3 out of 8 jobs. Inference with our scoped-learning model (depicted by the thin and bold rectangles) correctly labeled 7 out of 8 jobs.

estimate of its parameters, exactly supports the MAP approximation in Section 2.3.1. However, by treating the local parameter as a random variable, we can explicitly treat each locale as IID and integrate out ϕ. As we've seen in Section 4.3, this leads to a significant increase in performance over the MAP estimate.

Our work is also related to learning supervised tasks using a combination of labeled and unlabeled data—particularly *transduction* in which the unlabeled set is the test set (Joachims, 1999). The connection is that we have no truly labeled data for a particular locale and need to learn the local parameters with uncertain labels obtained from an existing classifier. The previous work in this area however, does not model locales, represent a difference between local and global features, or have the opportunity to use locales at training time to learn hyperparameters over local features.

A final related thread is Slattery and Mitchell (2000), in which the authors build a system that automatically adjusts a trained classifier to the regularities discovered in the data which it is classifying. This is similar to the scoped-learning model, but their method has no explicit notion of locale and is not designed for any features other than web page links. Furthermore, since it is not a probabilistic model, their method does not afford such benefits as integrating out parameter uncertainty.

6 Conclusion

We have presented a probabilistic model and training/inference procedures which can exploit both local and global features in various prediction tasks. On two large, real-world problems our model performs significantly better than the traditional global classifier, in some cases reducing error by more than half. In comparing our approximation methods, we find that the MAP estimate tends to overfit while integration via variational methods alleviates this issue. Furthermore, the discriminative approach can outperform the generative methods particularly when the local features are not independent.

Though we have described the model in terms of web data, we would like to emphasize that the co-occurence of global features and scope-limited features naturally arises in other domains besides text. Furthermore, we have developed a model based on only two levels of scope (local and global). It should be clear that this

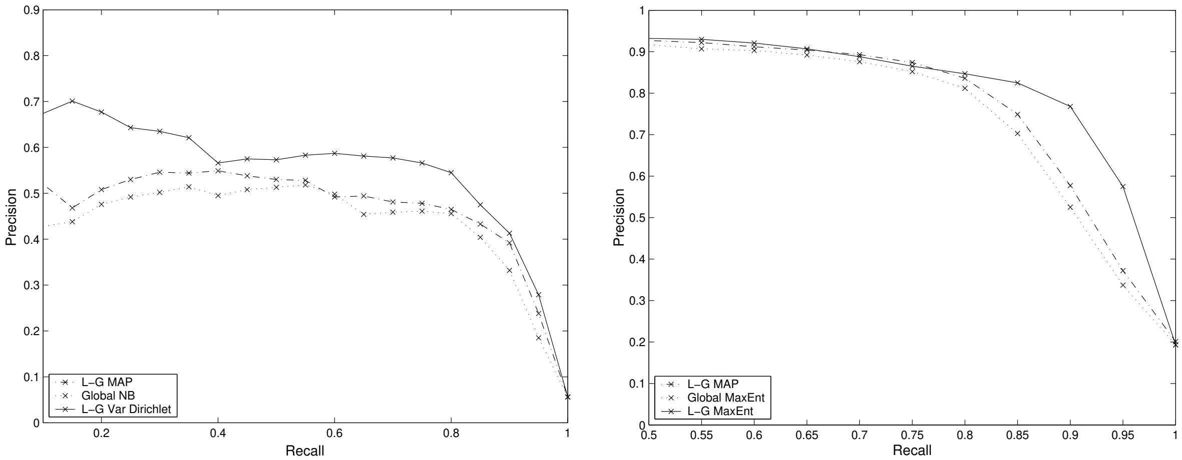

Figure 3: Scoped learning dominates the global classifier alone. (Left) A precision-recall curve that summarizes the the experimental results on the JOB-TITLE information extraction task. (Right) A precision-recall curve that summarizes the experimental results on the PRESS-RELEASE classification task.

concept can be recursively extended to arbitrary levels. For example some features may be common to all English language web-sites, some to a particular corporate domain, and some to particular web page.

Finally, we have developed and tested our model using simple mixture models for the underlying local and global models. It is quite natural to extend our work to incorporate more sophisticated models of local and global features as well as the relationship between class labels. For example, while we treated the IE task as one of independent classification, it is more natural to model it as a finite-state sequence problem. Extending our model this way is straightforward by introducing a chain-structure to the class labels. Under this model, our inference algorithms become a variant of standard inference procedures on hidden Markov models.

There is promising future work on this model of scoped learning. First, we assume that the local and global features are disjoint. One can imagine models in which the same features appear in both the global and local settings and inference must determine which model has generated a particular occurrence of a feature. This would be useful, for example, in situations where word usage exhibits local regularities. Second, we have assumed a uniform distribution over the local parameters for each locale. There may be situations in which one would want to find a maximum likelihood estimate of these parameters to determine how useful they will be in future data.

Acknowledgments

This work was done while all three authors were working at WhizBang Labs during the summer 2001, often discussing these matters while playing pool. We would like to thank Huan Chang for help with the job title data, William Cohen for advice and code related to generating local page features, Jonathan Baxter, Tom Minka, and Andrew Ng for insightful discussions, and Dallan Quass for inspirational work on extraction.

References

Attias, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*.

Blei, D., Ng, A., and Jordan, M. (2002). Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 14*.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Conference on Computational Learning Theory*.

Cohen, W. and Jensen, L. (2001). A structured wrapper induction system for extracting information from semi-structured documents. In *Proceedings of the IJCAI 2001 Workshop on Adaptive Text Extraction and Mining*.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proc. 16th International Conf. on Machine Learning*, pages 200–209. Morgan Kaufmann, San Francisco, CA.

Slattery, S. and Mitchell, T. (2000). Discovering test set regularities in relational domains. In *Proc. 17th International Conf. on Machine Learning*, pages 895–902. Morgan Kaufmann, San Francisco, CA.

Taskar, B., Segal, E., and Koller, D. (2001). Probabilistic classification and clustering in relational data. In *IJCAI*, pages 870–878.