A Hierarchical, Hierarchical Pitman Yor Process Language Model

Frank Wood FWOOD@GATSBY.UCL.AC.UK
YeeWhye Teh YWTEH@GATSBY.UCL.AC.UK
Gatsby Computational Neuroscience Unit, University College London, London, WC1N 3AR UK

Keywords: Hierarchical Dirichlet process, language modeling, domain adaptation

Abstract

In this paper we present a novel nonparametric Bayesian approach to domain adaptation for statistical language models. Specifically we describe a model consisting of a *hierarchy* of hierarchical Pitman-Yor language models (Teh, 2006; Goldwater et al., 2007), show one way to estimate such a model, and explain how inference in such a model can be intepreted as a kind of Bayesian interpolation between language models. We provide empirical evidence that this approach is sound by demonstrating improved smoothing between disparate corpora.

1. Introduction

There are many real-world domains for which one may not have a sufficient quantity of training data to reliably estimate a useful model. Obtaining sufficient quantities of training data for these "specific" domains can be a significant logistical or economic challenge. In some cases, however, there may already exist a large quantity of training data from a related or more "general" domain. The phrase domain adaptation is used to describe modeling techniques that seek to utilize general data to improve modeling of specific domains (Daumé III & Marcu, 2006; Ben-David et al., 2007).

Various approaches for solving this problem in language modeling have been proposed and studied. Mixtures of n-gram languge models in which the mixing weights themselves were a function of some number of preceding words were studied by (Kneser & Steinbiss, 1993). Static mixtures of smoothed n-gram models and various methods of estimating single n-gram models by combining count statistics from multiple corpora were compared by (Iyer et al., 1997). Using world wide web search engine query results to improve trigram models was found to improve n-gram language modeling by (Zhu & Rosenfeld, 2001). A recent and quite extensive review of these and several other adaptation approaches can be found in (Bellegarda, 2004).

Starting with a brief review of hierarchical Pitman Yor process (HPYP) language modeling, we then introduce of the hierarchy of hierarchical Pitman Yor process (HHPYP) language model which addresses the problem of domain adaptation in langauge modeling.

2. HPYP

In a normal Pitman Yor process language modeling, the distribution over words following a particular context (set of antecedent words)

$$w_t|w_{t-1}, w_{t-2} \sim \mathcal{G}^0_{\{w_{t-2},w_{t-1}\}}$$

(here the context length is two) is itself a random distribution

$$\mathcal{G}^0_{\{w_{t-2},w_{t-1}\}} \sim \mathcal{PY}(d_2, \alpha_2, \mathcal{H})$$

where $\mathcal{PY}(d, \alpha, \mathcal{H})$ is a Pitman Yor process with discount d, concentration α, and base distribution $\mathcal{H}$. When the base distribution is the distribution over words following the same context with one fewer antecedents, $\mathcal{H} = \mathcal{G}^0_{\{w_{t-1}\}}$, and $G^0_{\{w_{t-1}\}}$ is itself a random distribution which is distributed according to a Pitman Yor process with yet another more general base distribution, then the resulting model is referred to as a HPYP language model. This "recursion" continues until the set of antecedent words is empty; in that case the "root" Pitman Yor process is typically given a base distribution which is uniform over the corpus vocabularly. Such a HPYP language model is clearly very closely related to a hierarchical Dirichlet process and model estimation procedures such as Gibbs sampling

Preliminary work. Under review by the ICML Nonparametric Bayes Workshop 2008. Do not distribute.

can be straightforwardly adopted. The most helpful intuition about the HPYP language model comes from its relationship to non-Bayesian language model smoothing in which the distribution over words following a long context "backs-off to", or alternatively stated, is "centered on" a distribution over words following a shorter context. As training data counts grow sparse for word/context combinations when contexts are long, it makes sense to design models that "fail" gracefully, resorting to less complex but easier to estimate models when the training data is insufficient to estimate the more complex model.

We carry this same intuition over into our development of the hierarchy of hierachical Pitman Yor process (HHPYP) language model. Assume that we have corpora from two domains $\mathcal{D}_1, \mathcal{D}_2$. While for domain adaptation purposes we could simply train a single HPYP model on the union of the two corpora, or form a convex combination of two HPYP models trained individually on each corpus, we instead take the approach common to Bayesian domain adaptation approaches and specify a hierarchical model that allows statistical sharing between the models of each corpus. The model we propose has the same form as the HPYP except that the base distribution of every Pitman Yor process in the hierarchy is different, namely

$$\mathcal{G}^{\mathcal{D}_i}_{\{w_{t-2},w_{t-1}\}} \sim \mathrm{PY}(d_j, \theta_j, \pi\mathcal{G}^{\mathcal{D}_i}_{\{w_{t-1}\}} + (1-\pi)\mathcal{G}^{0}_{\{w_{t-2},w_{t-1}\}}) \quad (1)$$

This choice of base distribution has the following intuitive justification: the distribution over words in a particular context in a particular domain could reasonably either back off to a distribution over words given a shorter context in the same domain or a distribution over words given the whole context in a general domain. Here π is the parameter that controls how closely the base distribution is tied to the domain specific model or the general model.

We call the statistical entity described by Equation 1 a "graphical Pitman Yor process" and establish posterior sampling algorithms for such a model. There is a natural extension of the Chinese restaurant process for such models which we call the multi-floor Chinese restaurant process.

3. Discussion

Encouraging preliminary HHPYP domain adaptation results have been established for models of the Brown and AMI corpora (Kucera & Francis, 1967; Carletta, 2007). A baseline test corpus perplexity (twenty thousand words from the AMI corpus, disjoint from the training data) was computing using a single HPYP model of the union of a million word subset of the Brown corpus and a six hundred thousand word subset of the AMI corpus was established. An HHPYP model was trained using the same Brown and AMI corpora subsets. The HHPYP achieved lower test perplexity than the HPYP model of the same data. Additionally, our experiments suggest that using an HHPYP language model for domain adaptation may require less domain specific training data than a naive model to achieve a given test corpus perplexity.

Acknowledgments

This work was supported by the Gatsby Charitable Foundation.

References

Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech Communication*, *42*, 93–108.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. *NIPS 19* (pp. 137–144).

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, *41*, 181–190.

Daumé III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 101–126.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2007). Interpolating between types and tokens by estimating power law generators. *NIPS 19* (pp. 459–466).

Iyer, R., Ostendorf, M., & Gish, H. (1997). Using out-of-domain data to improve in-domain language models. *IEEE Signal processing letters*, *4*, 221–223.

Kneser, R., & Steinbiss, V. (1993). On the dynamic adaptation of stochastic language models. *IEEE Conference on Acoustics, Speech, and Signal Processing* (pp. 586–589).

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.

Rosenfeld, R. (2000). Two decages of statistical language modeling: where do we go from here? *Proceedings of the IEEE* (pp. 1270–1278).

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. *ACL Proceedings (44th)* (pp. 985–992).

Zhu, X., & Rosenfeld, R. (2001). Improving trigram language modeling with the world wide web. *IEEE Conference on Acoustics, Speech, and Signal Processing* (pp. 533–536).