A *Hierarchical* Hierarchical Pitman-Yor Process Language Model

Frank Wood, Yee Whye Teh {fwood,ywteh}@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, London.

The Problem

Statistical natural language model "domain adaptation" describes adapting a language model trained on a large general corpora to "fit" a specific domain for which less training data is available [1,3,5,6,8,10]. This is important to do because, while it may be possible to train even relatively large language models (e.g. using text from the world wide web), the resulting models are often ill-suited for specialized application domains. For instance the language used in a company's customer service interactions may differ significantly from the more general language found on the world wide web [1].

Our Approach

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 process language models [4,9] show one way to estimate such a model, and explain how inference in such a model can be interpreted as a kind of Bayesian interpolation between language models. We provide empirical evidence that this approach is sound by demonstrating improved modeling results for disparate corpora.

Review – HPYP Language Model (HPYLM)

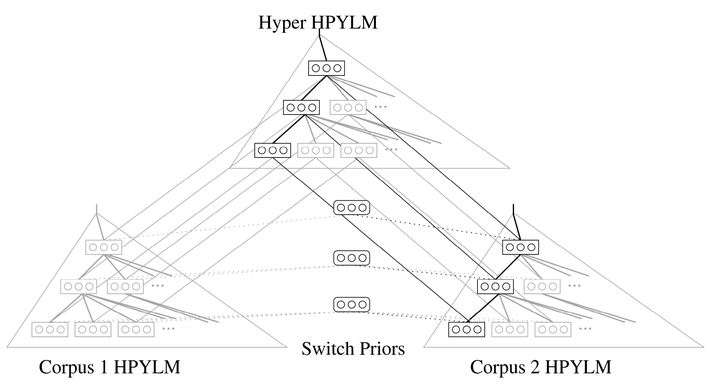
The HPYLM is a hierarchical Pitman-Yor process (HPYP) [9] which is itself closely related to the Hierarchical Dirichlet process.

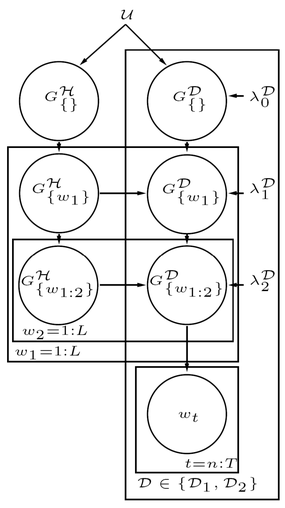
$$
\begin{aligned}
\mathcal{G}_{\{\}} &\sim \mathrm{PY}(d_0, \alpha_0, \mathcal{U}) \\
\mathcal{G}_{\{w_{t-1}\}} &\sim \mathrm{PY}(d_1, \alpha_1, \mathcal{G}_{\{\}}) \\
&\vdots \\
\mathcal{G}_{\{w_{t-n+1:t-1}\}} &\sim \mathrm{PY}(d_{n-1}, \alpha_{n-1}, \mathcal{G}_{\{w_{t-n+2:t-1}\}}) \\
w_t | w_{t-n+1:t-1} &\sim \mathcal{G}_{\{w_{t-n+1:t-1}\}}
\end{aligned}
$$

d discount
α concentration
w_t word
$\mathcal{G}$ distribution over finite vocabulary
$\mathcal{U}$ uniform distribution over vocabulary

Hierarchical HPLYM (HHPYLM)

The HHPYLM consists of a set of HPYLM's, one each for each domain, coupled under a shared "hyper" HPYLM. "Domain adaptation" is automatic in this model via hierarchical sharing of priors.





"Hyper" Language model

$$
\begin{aligned}
\mathcal{G}^{\mathcal{H}}_{\{\}} &\sim \mathrm{PY}(d^{\mathcal{H}}_0, \alpha^{\mathcal{H}}_0, \mathcal{U}) \\
\mathcal{G}^{\mathcal{H}}_{\{w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{H}}_1, \alpha^{\mathcal{H}}_1, \mathcal{G}^{\mathcal{H}}_{\{\}}) \\
&\vdots \\
\mathcal{G}^{\mathcal{H}}_{\{w_{t-j:t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{H}}_j, \alpha^{\mathcal{H}}_j, \mathcal{G}^{\mathcal{H}}_{\{w_{t-j+1:t-1}\}})
\end{aligned}
$$

$\{\lambda, (1\text{-}\lambda)\}$ context "switching" distribution

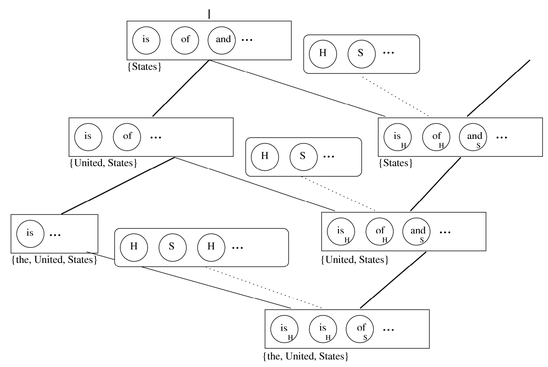
Domain specific language model(s)

$$
\begin{aligned}
\mathcal{G}^{\mathcal{D}}_{\{\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_0, \alpha^{\mathcal{D}}_0, \lambda^{\mathcal{D}}_0 \mathcal{U} + (1-\lambda^{\mathcal{D}}_0) G^{\mathcal{H}}_{\{\}}) \\
\mathcal{G}^{\mathcal{D}}_{\{w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_1, \alpha^{\mathcal{D}}_1, \lambda^{\mathcal{D}}_1 \mathcal{G}^{\mathcal{D}}_{\{\}} + (1-\lambda^{\mathcal{D}}_1) \mathcal{G}^{\mathcal{H}}_{\{w_{t-1}\}}) \\
&\vdots \\
\mathcal{G}^{\mathcal{D}}_{\{w_{t-j}:w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_j, \alpha^{\mathcal{D}}_j, \lambda^{\mathcal{D}}_j \mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1}:w_{t-1}\}} + (1-\lambda^{\mathcal{D}}_j) \mathcal{G}^{\mathcal{H}}_{\{w_{t-j}:w_{t-1}\}}) \\
w_t | w_{t-n+1} : w_{t-1} &\sim \mathcal{G}^{\mathcal{D}}_{\{w_{t-n+1}:w_{t-1}\}}.
\end{aligned}
$$

Estimation

A "multi-floor" Chinese restaurant franchise (MFCRF) representation is used to construct a Gibbs sampler for an equivalent model in which the $\mathcal{G}$'s are integrated out in the usual way. The MFCRF representation and sampler is the same as the Chinese restaurant franchise sampler for the HDP, except that each restaurants' tables have two labels, one the usual and the other a "floor" indicator variable. The floor variables indicate from which component of the base distribution the table came.

Intuition



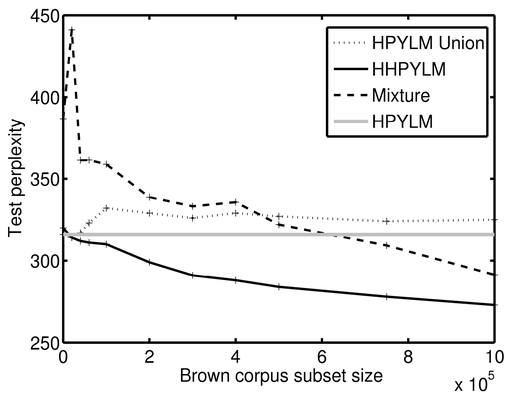
An Aside: The "Graphical Pitman-Yor Process"

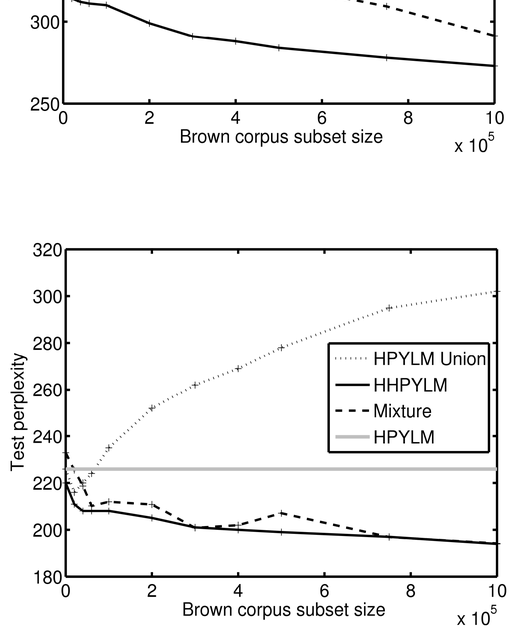
$$
\begin{aligned}
\Lambda_v &\sim \mathcal{S}_v \\
\mathcal{G}_v | \{G_w : w \in \mathrm{Pa}(v)\} &\sim \mathrm{PY}(d_v, \alpha_v, \sum_{w \in \mathrm{Pa}(v)} \lambda_{w \to v} \mathcal{G}_w)
\end{aligned}
$$

The HHPYLM is a graphical Pitman-Yor process. The conditional updates for the indicator variables in the MFCRF are given by

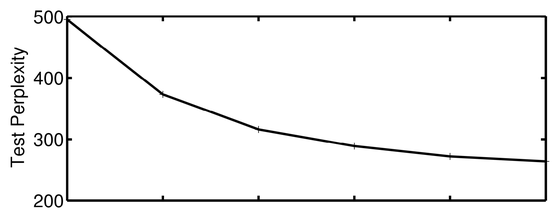
$$
\begin{aligned}
P(z^j_v = k | \mathbf{z}_v \backslash z^j_v, S, X) &\propto \max((c^{k-}_v - d_v), 0) \delta(x^j_v - \phi^k_v) \\
P(z^j_v = K+1, s^{K+1}_v = w | \mathbf{z}_v \backslash z^j_v, S, X) &\propto (\alpha_v + d_v K^-_v) \lambda_{w \to v} \mathcal{G}_w(x^j_v).
\end{aligned}
$$

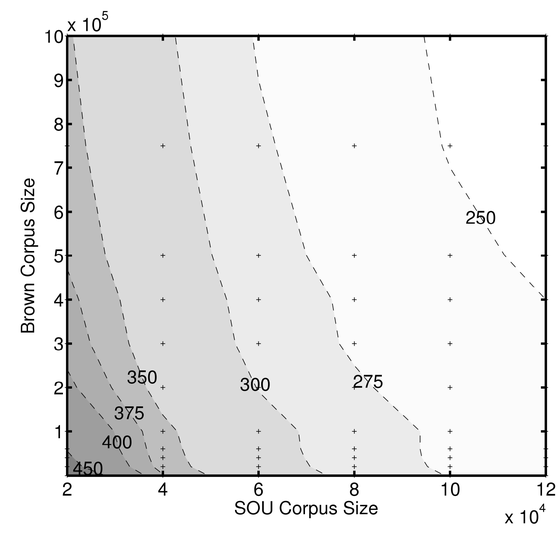
Experiments





Relative performance of various domain adaptation approaches for two different corpora (Top: State of the Union, Bottom: AMI [2]). Both graphs show test perplexity as a function of Brown training corpus [7] size. Lower test perplexity is better.





Top: HPYLM "baseline" test perplexity of Lyndon Johnson's state of the union addresses as a function of training corpus size. Bottom: test perplexity for the HHPYLM. Lower test perplexity is better. The entire lower graph is below the baseline. Both the HPYLM and HHPYLM were trained using other US presidents' state of the union addresses (x-axis; shared with the top plot). The HHPYLM used subsets of the Brown corpus (y-axis) for its out of domain, general corpus.

Acknowledgments

This work was supported by the Gatsby Charitable Foundation.

References

[1] Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. Speech Communication, 42, 93–108.
[2] Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. Language Resources and Evaluation Journal, 41, 181–190.
[3] Daum´e III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research, 101–126.
[4] Goldwater, S., Griffiths, T. L., & Johnson, M. (2007). Interpolating between types and tokens by estimating power law generators. NIPS 19 (pp. 459–466).
[5] 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.
[6] Kneser, R., & Steinbiss, V. (1993). On the dynamic adaptation of stochastic language models. IEEE Conference on Acoustics, Speech, and Signal Processing (pp. 586–589).
[7] Kucera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Brown University Press.
[8] Rosenfeld, R. (2000). Two decages of statistical language modeling: where do we go from here? Proceedings of the IEEE (pp. 1270–1278).
[9] Teh, Y.W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. ACL Proceedings (44th) (pp. 985–992).
[10] 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).