Scalable sequence memoization for natural language modeling and lossless compression.

PI	Frank Wood, Ph.D.	**Position**	Assistant Professor
Co-PI	David Madigan, Ph.D.	**Position**	Professor
Address	Room 1005 SSW, MC 4690 1255 Amsterdam Avenue New York, NY 10027	**University** **Department**	Columbia University Statistics
Phone	212.851.2132	**Fax**	212.851.2164
Websites	http://www.stat.columbia.edu/∼fwood http://www.stat.columbia.edu/∼madigan		

Abstract

We propose to develop and demonstrate scalable inference software for a Bayesian nonparametric (BNP) natural language model called the sequence memoizer (SM) [5]. We propose to use this software to train a SM using the trillion word Google text corpus and to empirically study the effect of taking into account long-range textual dependencies on language model performance as the amount of training data grows. We also propose to develop a latent variable extension of the SM and demonstrate scalable inference in the same. We propose to demonstrate both models and algorithms in two ways, general purpose lossless compression and n-gram natural language model performance.

Keywords : Bayesian nonparametric inference, n-gram natural language modeling, sequence memoizer
Research Areas : language modeling, machine translation, automated speech recognition, compression

Goals

- Develop a freely available, downloadable software development kit (SDK) that contains a scalable implementation of a constant space, linear time SM language model that demonstrably scales to sequences that are billions to trillions of tokens long.
- Empirically explore the impact of being able to probabilistically model and exploit long contextual dependencies given Google-scale corpora on language model perplexity and compressor log-loss.

Expected Outcomes

- The primary result of this project will be the establishment of compelling evidence that supports the practicality and usefulness of Bayesian nonparametric language models for large scale commercial applications including machine translation, automated speech detection, and general purpose lossless compression.
- We will develop a SM SDK that will be downloaded and used by researchers in a wide-variety of industrial and academic fields.
- We will contribute to the state of the art in Bayesian nonparametric modeling by developing an latent variable extension to the SM.

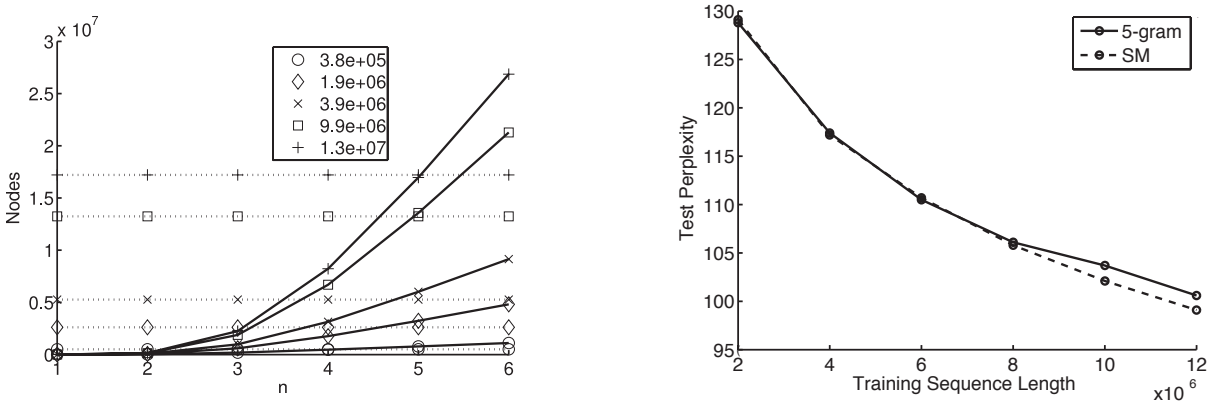
Big Picture

Traditional parametric statistical tools and methods are designed to allow inference about a population from a small sample. While this parametric style of inference will always have a place, it is now the case that one often has access to so much data that parametric models themselves are sometimes not even necessary. In other words, for some problems one can do strict nonparametric inference; i.e. one can query the data directly. While such a nonparametric approach has attractive characteristics and for some problems is feasible to consider, particularly given Google-scale data, we suggest that there are stochastic processes of sufficient complexity (e.g. natural language generators) that strictly nonparametric approaches to estimation and inference (e.g. non-smoothed n-gram language models) are bound to fail. To attack these kinds of problems, we advocate and actively pursue computationally practical ways to estimate and perform inference in Bayesian nonparametric (BNP) models. BNP models are nonparametric in nature, which gives them inferential capacity that can be understood to grow as a function of the amount of training data. Contrast this to parametric models in which inference is limited to inference through and about a finite set of parameters. In parametric models as the amount of data grows, posterior uncertainty about the value of the parameters will vanish given sufficient data, rendering the inclusion of additional data irrelevant. This

is not true for BNP models, making them ideal for inference in the modern regime of continually growing data. BNP models are are also Bayesian in nature which allows for hierarchical Bayesian-style regularization and incremental Bayesian-style inference and estimation. For small scale data on the order of millions of observations, BNP natural language models and lossless compressors we have recently been shown to exhibit excellent empirical characteristics [4, 5, 2]. Unfortunately BNP models in general have been saddled with an unfortunate stigma, namely that they are as a class uniformly computationally complex. We suspect that this stigma is at least partially responsible for holding back wide adoption of BNP methods. The work outlined in this proposal aims to chip away at this stigma by providing concrete evidence that at least one member of this class of BNP models scales well.

Scaling the Sequence Memoizer

In recent work we established a rather surprising result: for non-antagonistically generated discrete sequence data (natural language token sequences, bytes, bits, etc.), we found that it was possible to estimate the SM in the same asymptotic space and time as is required to estimate a smoothing 5-gram model [5] (Figure 1(a)). One way to understand the SM is as a Bayesian smoothing n-gram model in the limit of n taken to infinity. This means that the SM is capable of modeling long-range dependencies in discrete sequence data as opposed to finite-order Markov models which cannot. What is more, in related work, we uncovered empirical evidence that supports Shannon's assertion [3] that long range dependencies in written language do exist and are significant up to and potentially extend beyond hundreds of characters [2] (related evidence for this can also be seen in Figure 1(b)).



(a) (b)

Figure 1: (a) Memory complexity and estimation computational cost (in units of graphical model "nodes") versus n of n-gram for the SM and a smoothing n-gram model. There is one pair of lines for each of 5 different observation sequence lengths. The horizontal dotted lines are the SM computational requirements. The quadratic solid lines are the computational requirements the smoothing n-gram models. The SM always uses all contextual information when performing predictive inference. The n-gram model discards contextual observations more distant than $n-1$. Nonethe-less (a) shows that the computational complexity of each standard n-gram model exceeds that of a sequence memoizer trained on the same data for all $n > 5$ and for all training observation sequence lengths. This suggests that the sequence memoizer should be considered as an alternative to n-gram language modeling if long-range contextual dependencies matter. The data used in this experiment was an excerpt of the New York times corpus. (b) Held-out test perplexity for two language models trained on growing excerpts of the Associated Press news corpus. 5-gram is a hierarchical Pitman-Yor process language model (a generalization of a Kneser Ney smoothed 5-gram model). SM is the sequence memoizer. Training sequence length ranges from 2-12 million tokens which we believe to be too small to characterize the benefit of being able to model and use long range dependencies. Regardless, (b) suggests that held-out perplexity under the SM seems to improve relative to the smoothed n-gram model as more data is introduced into the model. We conjecture that this is due to increasing prevalence of meaningful long contexts. Experiments such as those proposed herein on larger corpora would establish the veracity of this conjecture. Both of these figures were taken from [5].

We have already begun to demonstrate the practical utility of the SM, and in particular the benefits it extends beyond fixed, finite-depth n-gram models. Various studies have found evidence of improvements to natural language models for automated speech recognition and machine translation, better encoding models for lossless compressors, and so forth. These benefits seem to accrue from by being able to model and use the extra information that longer

contexts provide, but much more experimentation is required to verify this observation. We feel that an important scientific question remains unanswered: what happens to the importance of long-range contextual information as the amount of training data is increased? Our intuition suggests that long-range dependencies will become more valuable, and the advantage of the SM relative to n-gram language models will grow, but currently the answer to this question remains unknown. Further, the software and hardware infrastructure needed to answer this question do not currently exist. One of the main purposes of this proposal is to start to address this question.

We have taken several theoretical and practical steps beyond the initial version of the SM; incremental estimation of the SM and the coupling of it to an entropy coder resulted in a highly competitive general purpose lossless compressor (significantly better than, for instance, gzip and bzip2) that scaled sufficiently to encode and compress a 100MB wikipedia corpus (1.6 bits / byte) [2]. Following on that work, we have taken steps towards the development of a constant space, linear time approach to estimation of a class of models that includes the SM [1]. This work opens up for the first the possibility of performing experiments on large scale sequence data including the tera-word Google corpus. No one knows what will happen when a powerful NPB language model such as the SM is trained on orders of magnitude more data than ever before. Most of the engineering work necessary to achieve the asymptotic complexity results established in [1] (constant space storage, linear time estimation, constant time inference) remains to be done, but we know it can be. This means that many of the most interesting questions about the value of long-range contextual dependencies and the inferential power of BNP language models finally have a chance to be answered.

Budget

Travel	2 International conferences	$5,000
	2 Domestic conferences	$3,000
Hardware	2 Laptops, one high memory server, compute cloud time	$20,000
Salary	Post-doctoral fellow (1 year)	$60,000
	PI summer (2 months)	$14,667
	Fringe 28.5%	$21,280
Total (salary)		$95,947
Total		$123,947

Justification

The PI is requesting summer salary that will allow him to concentrate his efforts on this research work. The Co-PI's request travel funding for themselves and one student each to travel to one international and one domestic conference in the upcoming year. The Co-PI's have also requested funds for a high-memory server and compute cloud time for the simulation studies described herein, and laptop computers to carry-out work while away from campus.

Google Contacts

The following Google researchers are personally familiar with one or more of the PI's on this proposal. No specific technical sponsor for this work has yet been selected.

Thomas Hoffman Tom Dean Daryl Pregibon Diane Lambert Steven Scott

References

[1] Bartlett, N., Pfau, D., and Wood, F. (2010). Forgetting counts : Constant memory inference for a dependent hierarchical Pitman-Yor process. In *(*submitted) ICML.

[2] Gasthaus, J., Wood, F., and Teh, Y. W. (2010). Lossless compression based on the Sequence Memoizer. In *Data Compression Conference 2010*, pages 337–345.

[3] Shannon, C. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, pages 50–64.

[4] Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the Association for Computational Linguistics*, pages 985–992.

[5] Wood, F., Archambeau, C., Gasthaus, J., James, L., and Teh, Y. W. (2009). A stochastic memoizer for sequence data. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1129–1136, Montreal, Canada.