Hierarchical Pitman-Yor Language Model for Information Retrieval

Saeedeh Momtazi, Dietrich Klakow
Spoken Language Systems
Saarland University, Saarbrücken, Germany
{saeedeh.momtazi,dietrich.klakow}@lsv.uni-saarland.de

ABSTRACT

In this paper, we propose a new application of Bayesian language model based on Pitman-Yor process for information retrieval. This model is a generalization of the Dirichlet distribution. The Pitman-Yor process creates a power-law distribution which is one of the statistical properties of word frequency in natural language. Our experiments on Robust04 indicate that this model improves the document retrieval performance compared to the commonly used Dirichlet prior and absolute discounting smoothing techniques.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]:Information Search and Retrieval

General Terms: Theory, Algorithm, Experimentation

Keywords: information retrieval, language modeling, Pitman-Yor process, smoothing methods

1. INTRODUCTION

Statistical language modeling has successfully been used in speech recognition and many natural language processing tasks. Language models for information retrieval have been the topics of intense research interest in recent years. The efficiency of this approach, its simplicity, the state-of-the-art performance it provides, and straightforward probabilistic interpretation are the most important factors which contribute to its popularity [3].

Smoothing plays an essential role when estimating a language model for retrieving relevant documents. A large number of smoothing methods have been proposed for language modeling; among them, three different techniques—namely Jelinek-Mercer, Bayesian smoothing with Dirichlet priors, and absolute discounting—have shown significant improvements in information retrieval performance [6].

A hierarchical Bayesian language model based on Pitman-Yor processes has been recently proposed by Teh [5]. This model which is a nonparametric generalization of the Dirichlet distribution [5] has been shown to produce results superior to the state-of-the-art smoothing methods. Hierarchical Pitman-Yor language model has also been applied in speech recognition task and improved the system performance significantly [1]. However, to the best knowledge of the authors this method has not been used for language model-based information retrieval.

In this work, we propose using the hierarchical Pitman-Yor language model for the document retrieval task, and compare this approach with the state-of-the-art smoothing methods widely studied for language model-based information retrieval.

2. METHOD

In language model-based document retrieval, $P(Q|d)$ is estimated by the probability of generating each query term:

$$P(Q|d) = \prod_{i=1...M} P(q_i|d) \tag{1}$$

where M is the number of terms in the query, q_i denotes the ith term of query $Q = \{q_1, q_2, ..., q_M\}$, and d is the document model. Therefore, the goal is to estimate $P(w|d)$ which can be simply calculated by the maximum likelihood estimation:

$$P_{ml}(w|d) = \frac{c(w,d)}{\sum_w c(w,d)} \tag{2}$$

However, having the problem of unseen words, we need to use a smoothing technique to give a non-zero probability to the unseen words. We hypothesized that Bayesian smoothing based on Pitman-Yor process can be used as a new approach to solve the zero probability problem in document retrieval.

Pitman-Yor process is a nonparametric Bayesian model which recursively placed as prior for predicting probabilities in language model. Considering $P(w|d)$ as the probability of word w given the observation of document d to be estimated, the Pitman-Yor process can be defined as:

$$P(w|d) \sim PY(\delta, \mu, P_{BG}(w)) \tag{3}$$

where δ is a discount parameter, μ is a strength parameter, and P_{BG} is the prior/background probability of word w before observing any document.

The procedure of drawing word probabilities from the Pitman-Yor process can be described using the "Chines restaurant" analogy. Imagine a Chinese restaurant with an infinite number of tables, each with an infinite number of seats. Customers, which correspond to word tokens, enter the restaurant and seat themselves at a table. Each customer can sit at an occupied table k with probability $\frac{c_k - \delta}{\mu + c.}$ where c_k is the number of customers already sitting there and $c. = \sum_k c_k$;

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

the customer can also sit at a new unoccupied table with probability $\frac{\mu + \delta t.}{\mu + c.}$ where $t.$ is the current number of occupied tables. It is necessary to mention that all customers that correspond to the same word type w can sit at different tables, in which t_w denotes the number of tables occupied by customers w.

One of the advantages of Pitman-Yor process is improving the Dirichlet prior by using a discounting parameter δ ($0 < \delta < 1$) deriving from absolute discounting method. Another key advantage of Pitman-Yor process is generating a power-law distribution in the language model, which is one of the statistical properties of word frequencies in natural language. This property, which is based on the scenario of rich-get-richer, implies that in the statistical property of word counts, words with low frequency have a high probability and words with high frequency occur with low probability. Benefiting from this idea in the document smoothing can help us to have different discounting value for each word based on the frequency of that word in the document.

Given the seating arrangement of customers as described above, the estimated probability of word w having the observation of document d is given by:

$$P_{\mu\delta}(w|d) = \frac{c(w,d) - \delta t_w + (\mu + \delta t.)P_{BG}(w)}{\sum_w c(w,d) + \mu} \qquad (4)$$

If we set the discounting parameter $\delta = 0$, then the model reduces to the Dirichlet process. If we set the strength parameter $\mu = 0$ and limit $t_w = 1$, then the model reverts to the absolute discounting method.

Although this formula is based on unigram model, the hierarchical behavior of the Pitman-Yor process allows us to use this model for higher level n-grams as well.

The most important and computationally expensive part of the above formula is calculating t_w for each word which should have a relation to the word count $c(w,d)$. Towards this end, we use the power-law discounting model proposed by Huang and Renals [2]:

$$\begin{cases} t_w = 0 & \text{if } c(w,d) = 0 \\ t_w = f(c(w,d)) = c^{\delta}(w,d) & \text{if } c(w,d) > 0 \end{cases} \qquad (5)$$

They showed that the above formula is a near optimum estimate for t_w, which can be obtained without a computationally expensive training procedure.

3. EXPERIMENTAL RESULTS

To evaluate our methods, we used TREC ad hoc testing collections from disk 4 and 5 minus CR which includes Financial Times (1991-1994) and Federal Register (1994) from disk 4 and Foreign Broadcast Information Service (1996) and Los Angeles Times (1989-1990) from disk 5. The total number of documents are 528,155.

We used Robust04 topics for our experiment such that topics 301-450 have been used as development set and topics 601-700 for test set. For each of the topics, the set of top 1000 documents retrieved by Indri [4] was selected and then the documents are ranked with LSVLM, the language modeling toolkit developed by our chair, in the second step.

Table 1 shows the results of our experiments in which Mean Average Precision (MAP) and Precision at 10 (P@10) serve as the primary metrics, and results are marked as significant* ($p < 0.05$), highly significant** ($p < 0.01$), or neither according to 2-tailed paired t-test. This table presents our main results evaluating the accuracy of Bayesian smoothing with Dirichlet prior, absolute discounting and our proposed Bayesian smoothing based on Pitman-Yor process.

Table 1: Retrieval results with different smoothings. Significant differences with absolute discounting and Dirichlet prior are marked by *a* and *d* respectively.

Model	**MAP**	**P@10**
Absolute Discounting	0.3138	0.4484
Dirichlet Prior	0.3147	0.4518
Pitman-Yor Process	0.3271^{**a}_{*d}	0.4657^{*a}
Pitman-Yor Process ($\mu = 0$)	0.3222^{**a}	0.4566

As shown by the tabulated results, the Pitman-Yor language model significantly outperforms both Dirichlet prior and absolute discounting. As mentioned, the major features of the Pitman-Yor process are generalizing Dirichlet prior and generating power-law distribution by having different discounting parameters for each word based on its frequency. We believe that the power-law distribution is the main contribution of the Pitman-Yor language model which causes such an improvement in retrieval performance. We also applied Pitman-Yor language model while setting $\mu = 0$; i.e. the model became more similar to absolute discounting, but it still creates power-law distribution by benefiting from t_w parameter. The results are presented in the last raw of the table. From the results we can see that although setting $\mu = 0$ decreases the performance, the reduction is not significant; and the simplified version of Pitman-Yor smoothing which only has one parameter still beat the other smoothing methods.

4. CONCLUDING REMARKS

We proposed a new smoothing method for language model-based document retrieval, named Bayesian smoothing based on Pitman-Yor process, and verified that this language model provides better performance than other state-of-the-art smoothing techniques. The key advantage of Pitman-Yor language model is generating a power-law word distribution, which is the primary reason for its superior performance.

Acknowledgments

Saeedeh Momtazi is funded by the German research foundation DFG through the International Research Training Group (IRTG 715) “Language Technology and Cognitive Systems”.

5. REFERENCES

[1] S. Huang and S. Renals. Hierarchical Pitman-Yor language models for ASR in meetings. In *Proceedings of IEEE ASRU International Conference*, pages 124–129, 2007.

[2] S. Huang and S. Renals. Power law discounting for n-gram language models. In *Proceedings of IEEE ICASSP International Conference*, 2010.

[3] J. Ponte and W. Croft. A language modeling approach to information retrieval. In *Proceedings of ACM SIGIR International Conference*, pages 275–281, 1998.

[4] T. Strohman, D. Metzler, H. Turtle, and W. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of International Conference on Intelligence Analysis*, 2005.

[5] Y. Teh. A hierarchical Bayesian language model based on Pitman-Yor process. In *Proceedings of ACL International Conference*, 2006.

[6] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of ACM SIGIR International Conference*, 2001.