UNIVERSITY
of
GLASGOW

Azzopardi, L. and Girolami, M. and Van Rijsbergen, K. (2003) Investigating the relationship between language model perplexity and IR precision-recall measures. In, *Annual ACM Conference on Research and Development in Information Retrieval, July 28 - August 1 2003,* pages pp. 369-370, Toronto, Canada.

http://eprints.gla.ac.uk/3442/

Glasgow ePrints Service
http://eprints.gla.ac.uk

Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures

Leif Azzopardi
School of ICT
University of Paisley

Mark Girolami
School of ICT
University of Paisley

Keith Van Rijsbergen
Dept. of Computer Science
University of Glasgow

ABSTRACT

An empirical study has been conducted investigating the relationship between the performance of a generative language model in terms of perplexity and the corresponding information retrieval performance obtained. It is observed, on the corpora considered, that the perplexity of the language model has a systematic relationship with the achievable precision recall performance.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language Models*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Language modelling (LM) as an Information Retrieval (IR) paradigm has received much attention within the IR research community since Ponte and Croft [6] proposed LM for IR purposes. Others have continued the investigation of developing LM as a theoretically principled and competitive approach to IR [7, 4]. The application of statistical LM in IR is not new, the 2-Poisson and subsequent n-Poisson model [5] was one of the early attempts at the use of LM in IR. However, whilst an n-Poisson model obtains a good fit to the observed data (document collection) in terms of predictive likelihood or perplexity, a corresponding improvement in precision and recall (P-R) is not obtained. The canonical measure of goodness of a statistical LM is normally reported in terms of perplexity: the exponential of the negative normalized predictive likelihood under the model, and gives an indication of the expected word error rate as in speech recognizers. A fundamental question, which has not yet been considered, is that given two language models, one of which records lower perplexity than the other, what IR performance can be expected when employing these models. In other words, does lower model perplexity indicate superior P-R performance? This paper presents preliminary results which attempt to provide an answer to this question.

A latent variable unigram based LM, which has been successful when applied to IR, is the so called probabilistic latent semantic indexing (PLSI)[3]. The PLSI model has significantly reduced perplexity over smoothed unigram models and reports P-R figures that are superior to *cosine tf-idf* and latent semantic indexing (LSI)[2]. PLSI is in fact a *maximum a posteriori* (MAP) estimate of a latent Dirichlet allocation (LDA) model [1] under a uniform Dirichlet distribution; further elaboration of this point is not the focus of this paper. However, it provides a theoretical basis for the 'folding-in' of queries for PLSI where the procedure can be seen as obtaining the MAP estimate of the Dirichlet variables for the query. We employ PLSI in the following experiments to study the relationship between language model perplexity and IR precision-recall.

2. EXPERIMENTS

As a preliminary attempt at answering the question of whether LM perplexity is an indicator of expected P-R performance in our experiments we have taken four small collections as used in [3], the number of documents (d) and terms (t) follow the name of the corpus (Medline d=1033, t=5134; Cranfield d=1499, t=3262; CACM d=3204, t=4326; CISI d=1460, t=3918). Each collection was indexed by removing standard stop words, applying Porter Stemming and then removing any infrequent terms (those that appear only two times or less in the collection). Ten percent of each collection was randomly sampled to produce an out-of-sample set on which to measure perplexity[1] and the remainder was used to generate the model and perform indexing. This was performed ten times to produce ten different random sets for testing and parameter estimation. The model was then estimated with each collection and parameter estimation was ceased when the in-sample likelihood stopped increasing or any increase was negligible. Ten different random initializations of the parameters were used for each set. For each collection, and each model (k) a total of one hundred runs were generated. The number of latent variables in each model (k) ranged through [2, 4, 8, 16, 24, 32, 48, 64, 80, 96, 112, 128]. The methods for subsequent model based indexing as proposed in [3] were employed i.e. (PLSI-U and PLSI-Q).

[1]Perplexity was computed as detailed in [1] which is unlike

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2002 ACM X-XXXXX-XX-X/XX/XX ...$5.00.

Table 1: Best Performance Results

	MED	CACM	CISI	CRAN
Cos TF	43.7%	20.9%	17.3%	26.3%
Cos TF-IDF	51.2%	28.9%	17.0%	29.3%
PLSI-U	54.9%	29.2%	17.7%	29.8%
PLSI-Q	55.7%	31.6%	18.9%	30.7%

3. DISCUSSION AND CONCLUSIONS

As in [3] we provide the best results in terms of average precision obtained over all runs in Table. 1, which confirms that an increase in performance through the use of such models is achievable. To determine the actual significance of these results would require a study over all initializations and models. We omit such analysis due to space constraints and focus on the relationship between average precision and perplexity. In Figures 1 and 2, we show the average precision (AP) values for PLSI-U. The AP for each model (k) over 100 runs is represented by the mean and the error-bars signify one standard error (denoted by the $\diamond$ glyphs with values displayed on the right hand axis). The corresponding average perplexity is denoted by the $\square$ glyphs and the left hand axis displays the perplexity scores. The number of latent variables in the model (k) is shown along the bottom axis using a log scale.

We observe from the graphs a systematic relationship between perplexity and average precision with the exception of the CACM collection which seems to be resistant to reductions in perplexity. A one-way ANOVA performed on the results shows the differences in means of both perplexity and AP to be significantly different at the 5% level for all collections. However, follow-up Least Significant Difference (LSD) tests show that only a small number of the models actually exhibit a statistically significant difference in mean AP values. This can be observed visually by the scale of the AP values in relation to the error bars. What we can infer from these tests is that the effort expended in obtaining a low perplexity LM may not necessarily translate into significantly superior AP values. Interestingly, best performance using the latent variable style approach is gained when the model order is relatively low, between 8 and 64; though this may vary depending on the type of optimization method adopted. These results have been obtained using relatively small homogeneous corpora in IR terms. However, these findings should be taken seriously and warrant further investigation on larger and more heterogenous document collections. In addition other forms of language models such as those proposed in [6, 7, 4] would also complement these initial findings to further understand the nature of the relationship between LM perplexity and corresponding IR measures across different types of language models other than a latent variable language model.

In summary, this finding is important in that the focus of attention within the language modelling and machine learning community has been to devise low perplexity representations whilst the IR community has focused on employing LM to obtain high P-R performance. This work has examined language models for IR from both perspectives and the initial indications are that superior predictive LM's may not necessarily provide superior IR performance.

the original method of computing perplexity in [3].

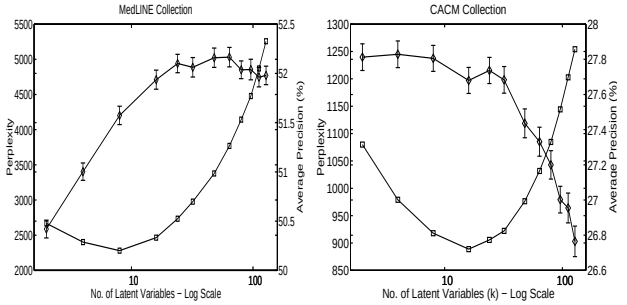

Figure 1: P-R and perplexity versus number of latent variables for MEDLINE and CACM.

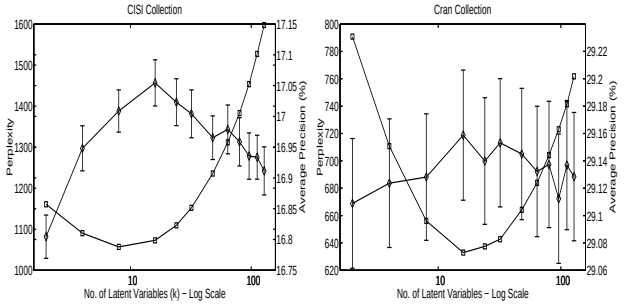

Figure 2: P-R and perplexity versus number of latent variables for CISI and CRAN.

4. ACKNOWLEDGMENTS

The first author is supported by Memex Technology Ltd.

5. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[3] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[4] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Research and Development in Information Retrieval*, pages 120–127, 2001.

[5] E. L. Margulis. Modeling documents with multiple poisson distributions. *Information Processing and Management*, 29(2):215–227, 1993.

[6] J. Ponte and W. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR 98*, pages 275–281. SIGIR, 1998.

[7] F. Song and W. B. Croft. A general language model for information retrieval (poster abstract). In *Research and Development in Information Retrieval*, pages 279–280, 1999.