Is it the Language Model in Language Modeling?

Warren R. Greiff
The MITRE Corporation
202 Burlington Road, Bedford, MA
greiff@mitre.org

This position paper explores the question of whether or not it is the use of language models, per se, that accounts for the recent surge of interest in what has come to be called Language Modeling in Information Retrieval (LMIR). We conjecture that, for the most part the answer is no. We suggest instead that the principal contribution of Language Modeling is that it makes patent the following:

1. the use of term frequencies in document evaluation can advantageously be viewed as statistical parameter estimation; and,
2. in so doing, Language Modeling (LM) approaches, explicitly or implicitly, address the role of variance reduction in producing models that result in improved retrieval performance.

We further suggest that recognition of the importance of estimation variance will have a beneficial effect on the continued development of theoretical foundations for Information Retrieval.

With the objective of supporting a more precise formulation of this question, and the discussion of the relevant issues, we begin with a formal definition of "language model". We then propose a description of what "the Language Modeling approach" can be thought to consist of, in the context of IR research; first with a strict interpretation in mind, and then with a more informal view. We conclude with a brief exposition of research into the role of variance reduction in IR recently begun at The MITRE Corporation.

What is Language Modeling in IR? - A Strict Interpretation

A language model is a probability distribution over a set of strings. More formally, given a vocabulary, V, the set of strings over V is given by:

$$v^* = \{\sigma \mid \text{for some } n;\ v_1, \ldots, v_n \in V : \sigma =< v_1, \ldots, v_n >\},$$

and a probability distribution over V^* is a mapping:

$$p : V^* \to [0, 1] \qquad \text{such that:} \quad \sum_{\sigma in V^*} p(\sigma) = 1.$$

It would be reasonable then to understand a "Language Modeling" approach to Information Retrieval as any approach to IR for which:

1. a language model is estimated;

2. the probability distributions of (one or more) language models enter into the calculations that are used to compare two arbitrary documents, relative to a specified information need.

If we restrict our understanding of language modeling to a definition such as this, the following would, strictly speaking, not then be sufficient for an approach to be considered an LM approach:

1. probabilistic modeling;
2. explicit recognition of the role of estimation;
3. document scoring in terms of. $p(q|d)$.

We note also that, under this view, neither of the last two would strictly be necessary for an approach to be considered Language Modeling.

What is Language Modeling in IR – Really?

While it will be useful to keep a strict interpretation of language modeling in mind, it is important as well to attempt to identify what it is, more informally, that results in certain research directions being understood as LM approaches. We propose that IR Language Modeling approaches share some combination of the following characteristics:

1. a language model is estimated, and plays an essential role in the assignment of Retrieval Status Values (RSV's);
2. document scoring is in terms of, $p(q|d)$;
3. parallels are drawn to, and ideas are adapted from, Language Modeling as a paradigm in other areas of human language technology;
4. the role of estimation is recognized.

Estimation of a language models: Starting from the top, we return to our claim that, no, "it is not the language models in Language Modeling". Language models are by no means new. There is a long history of taking term frequency (*tf*) as the probability of a word appearing in a document. Typically, a Poisson distribution is assumed, the two-Poisson model [6, 7] being, perhaps, the best known. Treatment of term frequency as a probability of word occurrence is not what is new. It is conscious attention to *tf* as the manifestation of an underlying probability distribution, rather than *tf* as the probability itself, and the concomitant concern for the question of estimation, that distinguishes recent work on Language Modeling from earlier research.

Probability of the query: Much of the LMIR work views query production as a stochastic process and takes the stand that ranking should be based on the probability that the given query would have been produced, conditioned in some way on the document. This represents a departure from the classical probabilistic perspective. Traditional probabilistic approaches adhered directly to the Probability Ranking Principle, which counsels ranking by the probability that documents will be found to be relevant, conditioned on the (fixed) query. An undercurrent in this paper is the belief that the scoring of documents by $p(q|d)$ instead of $p(d|q)$ does not account for the success of these models. Success, we assert, is due to variance reduction. For now, we leave as an open question exactly how this position may be formally expressed as a falsifiable hypothesis which can be experimentally tested.

Fertile metaphors: It is the third point that, perhaps more than others, gives a unifying theme for research into the use of Language Modeling for Information Retrieval. In his seminal work, Ponte discusses the influence that language modeling in other fields had had on the approach he developed for IR [8]. The application of Hidden Markov Models to Information Retrieval is clearly motivated by extensive use of this framework in Speech Recognition [10], and numerous areas of textual language processing, including part-of-speech tagging [4], named-entity identification [2], topic segmentation [5, 13], and selectional preference [1]. The value of adapting techniques developed, and leveraging the experience garnered, in these related areas of research should not be understated. Without minimizing the potential value of such cross-fertilization, we do not believe that, to date, this is what has been the primary contributor to the success of LM. That said, we focus our attention on the fourth, and final, point above.

Estimation and variance reduction: This paper makes the claim that it is recognition of parameter estimation as a fundamental issue in IR modeling that should be seen as the significant contribution of LMIR. Further, we believe it is the reduced variance of estimators used in LM approaches that accounts for the positive results that have been obtained. It is known that simply "shrinking" an estimate toward an arbitrary value can reduce mean squared error by trading bias for variance. A more informed choice for the shrinkage target can produce further improvement. Of course, reduced MSE does not translate automatically to improved retrieval performance. This relation must be studied. We believe these issues can be investigated and that such investigation will prove to be fruitful. We believe, also, that an appreciation for and understanding of the role of variance reduction will serve to place IR research on a sounder theoretical foundation, and that a sound theoretical underpinning will be essential if IR is to meet the expanding challenges that face it, as demands on information access technology increase.

Proposed Experimentation

In this final section we discuss three sets of experiments whose objective is to explore the relation of variance reduction and retrieval performance. Development of the simulation environment discussed in the next subsection is currently under way at The MITRE Corporation. Experimentation using this simulated environment, as well as the experimentation proposed in the following two subsections is planned for the near future. All of this work is motivated by interest in characterizing the relation between estimation variance and retrieval performance, and demonstrating the importance of variance reduction.

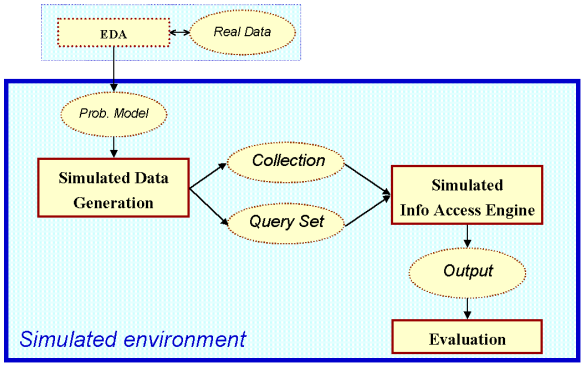

Figure 1: **Simulation experiments**

Simulation studies: In these studies, simulation will be used to create an idealized environment in which questions concerning estimation can be studied. Simulation allows for experimentation for which: confounding issues can be abstracted away; true probability distributions are known; and both environment and retrieval engine parameters can be controlled. Figure 1 gives a schematic view of the proposed simulation.

For initial simulation runs, a conditional log-linear model will be used to generate relevance judgments for the pseudo document collection. The log-odds of relevance will be given as a linear function of a logarithmic-like transform of the term frequency. The retrieval engine will also assume a log-linear model. However noise will be added to the true coefficients (those used to determine relevance judgments for the pseudo-collection) in order to produce the estimated coefficients used by the simulated engine. By introducing fluctuating additive noise with non-zero expected value, bias and variance can be introduced in a controlled manner. At first, term-frequency distributions and coefficients of the log-linear model will be designed in accordance with commonly held intuitions concerning retrieval situations based on years of IR research and current retrieval practice. In later stages of the study a greater effort will be made to have the setting of these simulation variables be informed by distributions extracted from existing test data, such as that provided by the TREC competitions [12]. This is shown pictorially in the upper left corner of Figure 1.

Variance reduction and traditional formulae: If estimation variance contributes to reduced retrieval performance, then any retrieval formula based on term frequencies should benefit from estimation procedures that reduce this variance. For example, the often used Okapi formula [11], is a probabilistically motivated formula. It is based on the two-Poisson model, but does not consider the introduction of bias in order to reduce variance as part of the estimation procedure. In this study, we will experiment with shrinkage estimators such as that used in [9], and study the effect on performance when used in conjunction with the Okapi formula. We also plan to experiment with other probabilistic formulae that have been used for retrieval. Finally, we will also look at the classical cosine similarity metric. While this is not a probabilistic approach, we believe that application of variance reduction techniques will have beneficial effects, nonetheless.

Alternate variance reducing estimators: In [8], Ponte uses a shrinkage estimator based on the geometric distribution. This is motivated by an interest in reducing the Bayesian Risk. Ponte did initiate an empirical study of alternatives to this smoothing mechanism. We plan to further pursue this line of research with the investigation of alternate estimators and their impact on retrieval performance, in the context of the Ponte/Croft Language Modeling approach. In particular, we plan to use Empirical Bayes methods [3] to exploit information given by the background distribution of term frequencies, in conjunction with information extracted from the document, in a principled way.

References

[1] Abney, S., and Light, M. Hiding a semantic hierarchy in a Markov model. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing, ACL* (1999).

[2] Burger, J. D., Palmer, D., and Hirschman, L. Named entity scoring for speech input. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics* (San Francisco, California, 1998), C. Boitet and P. Whitelock, Eds., Morgan Kaufmann Publishers, pp. 201–205.

[3] Carlin, B. P., and Louis, T. A. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London, 1996.

[4] Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing* (1992).

[5] Greiff, W., Morgan, A., Fish, R., Richards, M., and Kundu, A. Fine-grained Hidden Markov Modeling for broadcast-news story segmentation. In *Proceedings of the Human Language Technology Conference* (San Diego, Ca., March 2001).

[6] Harter, S. P. A probabilistic approach to automatic keyword indexing, Part I: On the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science 26* (1975), 197–206.

[7] Harter, S. P. A probabilistic approach to automatic keyword indexing, Part II: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science 26* (1975), 280–289.

[8] Ponte, J. M. *Probabilistic Language Models for Topic Segmentation and Information Retrieval*. PhD thesis, University of Massachusetts, Amherst, Massachusetts, May 1998.

[9] Ponte, J. M., and Croft, W. B. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia, Aug. 1998), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds., ACM Press, pp. 275–281.

[10] Rabiner, L. R., Levinson, S. E., and Sondhi, M. M. On the application of vector quantization quantization and hidden markov models to speaker-independent, isolated word recognition. *The Bell System Technical Journal 62*, 4 (April 1983), 1075–1106.

[11] Robertson, S. E. The probability ranking principle in IR. *Journal of Documentation 33* (1977), 294–304.

[12] Voorhees, E. M., and Harman, D. K. Overview of the eighth Text REtrieval Conference (TREC-8). In *The Eighth Text REtreival Conference (TREC-8)* (Gaithersburg, Md., 2000), E. M. Voorhees and D. K. Harman, Eds., NIST Special Publication 500-246, pp. 1–24.

[13] Yamron, J. P., Carp, I., Gillick, L., Lowe, S., and van Mulbregt, P. A Hidden Markov Model approach to text segmentation and event tracking. In *Proceedings ICASSP-98* (Seattle, WA., May 1998).