A General Language Model for Information Retrieval

Fei Song
Dept. of Computing and Info. Science
University of Guelph
Guelph, Ontario, Canada N1G 2W1
fsong@uoguelph.ca

W. Bruce Croft
Dept. of Computer Science
University of Massachusetts
Amherst, Massachusetts 01003
croft@cs.umass.edu

Statistical language modeling has been successfully used for speech recognition, part-of-speech tagging, and syntactic parsing [1]. Recently, it has also been applied to information retrieval ([5], [2], [3]). According to this new paradigm, each document is viewed as a language sample, and a query as a generation process. The retrieved documents are ranked based on the probabilities of producing a query from the corresponding language models of these documents.

One obstacle in applying statistical language modeling to information retrieval is the sparse data problem. It is serious in information retrieval for two reasons. First, a document is often too small, implying that many terms would be missing. If such a term were used in a query, we would get zero probability for the entire query. Second, a document is fixed in size and content, making it difficult to distinguish the effects of different missing terms in a document.

Our solution is to propose a new language model for information retrieval based on a range of data smoothing techniques. First of all, we smooth each document model with the Good-Turing estimate, which allocates some probability mass to the missing terms so that the probability of a missing term is always greater than zero [4]. Secondly, we expand each document model with the corpus model, with the intention of differentiating the missing terms. For example, in a corpus about information retrieval, the term "keyword" will likely to happen more often than the term "crocodile", so such information can be used to adjust the probabilities for the missing terms. Thirdly, we treat a query as a sequence of terms instead of a set of terms. One reason is that we want to handle the duplicate terms in a query. Of course, one can introduce weights into the set treatment of a query, but that will complicate the computation. The other reason is that we want to model phrases with local contexts, and this can only be done by viewing a query as a sequence of terms. Finally, we expand our term-based language model with the pair-based model. The intuition is that phrases such as term pairs would be useful in information retrieval, but the existing research often did not show much improvement in the retrieval performance. We would like to see that in the context of language modeling whether term pairs would bring any better results.

To demonstrate the effectiveness of our language model, we conducted experiments on two test collections. The Wall Street Journal (WSJ) data is a medium-sized homogeneous collection, with over 250 megabytes of information and 74,520 documents. The TREC4 data is a large heterogeneous collection, with over 2 gigabytes of information and 567,529 documents. Four different retrieval systems are used in our experiments. The baseline is the INQUERY system, and for the purpose of comparison, we also implemented Ponte and Croft's language model (LM). GLM(40) is our term-based general language model, where the combination between a document model and the corpus model is handled through interpolation (weighted sum), with the weighting parameter set to be 40% for the document model. GLM2(40+90) is our combined model for terms and term-pairs, with the weighting parameter between a document model and the corpus model set to be 40%, and the weighting parameter for the pair-based model set to be 90%.

As shown in table 1, the results of all the language models are comparable to that of INQUERY. In addition, our term-based model did 8.44% better than Ponte and Croft's model and our combined term-based and pair-based model did 16.38% better than Ponte and Croft's model. This is a clear indication that phrases of word pairs can be useful in improving the retrieval performance in the context of language modeling.

Table 1. Experimental Results on the WSJ Data Set

Retrieval Methods	11-pt Avg	%Change	%Change
INQUERY	0.2172	-	
LM	0.2027	- 6.68%	-
GLM(40)	0.2198	1.20%	8.44%
GLM2(40+90)	0.2359	8.61%	16.38%

For the large TREC4 data set, the results of the language models are once again comparable to that of INQUERY, as shown in table 2. However, the improvement of our models over Ponte and Croft's model is not as significant as that for the WSJ data set[1]. This is probably due to the heterogeneous nature of the TREC4 collection. In Ponte and Croft's model, there is a pathological problem in using the corpus probabilities: missing terms with a stopword feature are often assigned with high values, which could be problematic for a homogeneous collection, but less serious for

[1] In [5], Ponte and Bruce reported a significant improvement of their language model over INQUERY for the TREC4 data set. This is, however, not observed in our experiments. One reason for the difference may be the variation in preprocessing the raw TREC4 data [Ponte, personal communication].

a heterogeneous collection. Nevertheless, our models still did better than Ponte and Croft's and the term pairs are still shown to be useful in improving the retrieval performance. Note that our model has much potential for further improvement, since all the combination parameters can be individualized and optimized instead of setting them to be the same for all the documents.

Table 2. Experimental Results on the TREC4 Data Set

Retrieval Methods	11-pt Avg	%Change	%Change
INQUERY	0.1917	-	
LM	0.1890	- 1.41%	-
GLM(40)	0.1905	- 0.63%	0.79%
GLM2(40+90)	0.1923	0.31%	1.75%

In summary, we have proposed a simple yet intuitive language model for information retrieval. We also conducted experiments on two test collections. The results showed that the performance of our model is comparable to that of INQUERY and better than that of Ponte and Croft's language model. In particular, word pairs are shown to be useful in improving the retrieval performance. Our model can be potentially improved by individualizing and optimizing the combination parameters. Furthermore, our model is rooted on the solid foundation of statistical natural language processing. Any new techniques developed for data smoothing can be easily incorporated into our model. In this sense, the model serves as a general framework for language-based information retrieval.

ACKNOWLEDGMENTS

The authors would like to thank Jay Ponte, Stephen Harding, Margaret Connell, Mark Sanderson, Junxi Xu, Jeremy Pickens, and Lisa Ballesteros for their help and support.

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Charniak, E. Statistical Language Learning. The MIT Press, Cambrige MA, 1993.

[2] Hiemstra, D. A Linguistically Motivated Probabilistic Model of Information Retrieval. In Proceedings of the Second European Conference on Digital Libraries (1998), 569-584.

[3] Leek, T., Miller, D.R.H., and Schwartz, R.M. A Hidden Markov Model Information Retrieval System. In TREC-7 Proceedings (1998).

[4] Manning, C., and Schütze, H. Foundations of Statistical Natural Language Processing. Draft of August 31, 1998. To be published by the MIT press.

[5] Ponte, J.M., and Croft, W.B. A Language Modeling Approach to Information Retrieval. In Proceedings of SIGIR (1998), 275-281.