Statistical Language Models for Information Retrieval

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana_Champaign

Statistical language models have recently been successfully applied to many information retrieval problems. A great deal of recent work has shown that statistical language models not only achieve superior empirical performance, but also facilitate parameter tuning and provide a more principled way, in general, for modeling various kinds of complex and non-traditional retrieval problems.

The purpose of this tutorial is to systematically review the recent progress in applying statistical language models to information retrieval with an emphasis on the underlying principles and framework, empirically effective language models, and language models developed for non-traditional retrieval tasks. Tutorial attendees can expect to learn the major principles and methods of applying statistical language models to information retrieval, the outstanding problems in this area, as well as obtain comprehensive pointers to the research literature. The tutorial should appeal to both people working on information retrieval with an interest in applying more advanced language models and those who have a background on statistical language models and wish to apply them to information retrieval. Attendees will be assumed to know basic probability and statistics.

The outline of the tutorial is as follows:

1. Introduction
   (a) Information Retrieval (IR)
   (b) Statistical Language Models (SLMs)
   (c) Applications of SLMs to IR
2. The Basic Language Modeling Approach
   (a) Query likelihood document ranking
   (b) Smoothing of language models
   (c) Why does it work?
   (d) Variants of the basic LM
3. More Advanced Language Models
   (a) Improving the basic LM approach
   (b) Feedback and alternative ways of using LMs
4. Language Models for Special Retrieval Tasks
   (a) Cross-language IR
   (b) Distributed IR

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL pages 3–4,
Rochester, April 2007. ©2007 Association for Computational Linguistics

(c) Structured document retrieval
(d) Personalized/context-sensitive retrieval
(e) Expert retrieval
(f) Modeling redundancy
(g) Predicting query difficulty
(h) Subtopic retrieval

5. A General Framework for Applying SLMs to IR
    (a) Risk minimization framework
    (b) Special cases
    (c) Generative relevance hypothesis

6. Summary
    (a) SLMs vs. traditional methods: Pros & Cons
    (b) What we have achieved so far
    (c) Challenges and future directions

ChengXiang Zhai is an Assistant Professor of Computer Science at the University of Illinois at Urbana-Champaign, where he also holds a joint appointment at the Institute for Genomic Biology and the Graduate School of Library and Information Science. He received a Ph.D. in Computer Science from Nanjing University in 1990, and a Ph.D. in Language and Information Technologies from Carnegie Mellon University in 2002. He worked at Clairvoyance Corp. as a Research Scientist and, later, a Senior Research Scientist from 1997 to 2000. His research interests include information retrieval, text mining, natural language processing, machine learning, and bioinformatics. He serves on the editorial board of ACM Transactions on Information Systems, and is the program co-chair of ACM CIKM 2004 and NAACL HLT 2007. He is an invited participant of the National Academy of Engineering's 2006 US Frontiers of Engineering Symposium. He received an NSF CAREER Award in 2004, the ACM SIGIR 2004 Best Paper Award, and the 2004 Presidential Early Career Award for Scientists and Engineers (PECASE).