

6. Statistical Language Models and Information Retrieval

ChengXiang Zhai, University of Illinois

<http://www-faculty.cs.uiuc.edu/~czhai>

Statistical language models play an important role in virtually all kinds of tasks involving human language technologies. In particular, they have been attracting much attention recently in the information retrieval community due to their theoretical and empirical advantages over traditional retrieval methods. A great deal of recent work has shown that statistical language models not only lead to superior empirical performance, but also facilitate parameter tuning, open up possibilities for modeling non-traditional retrieval problems, and in general provide a more principled way of modeling 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.

6.1. Tutorial outline

1. Introduction
 1. Information Retrieval (IR)
 2. Statistical Language Models (SLMs)
 3. Applications of SLMs to IR
2. The Basic Language Modeling Approach
 1. Query likelihood methods and their justification
 2. Smoothing of language models
 3. Improving the basic language modeling approach
3. Feedback Language Models
 1. Different ways of feedback with language models
 2. Representative feedback models (relevance/query models, translation models)
4. Language Models for different retrieval tasks
 1. Cross-language retrieval
 2. Distributed information retrieval
 3. TDT and information filtering
 4. Semi-structured information retrieval
 5. Subtopic retrieval
5. A General Framework for Applying SLMs to IR
6. Summary
 1. SLMs vs. traditional methods: Pros & Cons
 2. Progress so far
 3. Challenges and future research directions

6.2. Target Audience

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.

ChengXiang Zhai is an Assistant Professor of Computer Science at the University of Illinois at Urbana-Champaign. 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 broadly include information retrieval, natural language processing, machine learning, and bioinformatics. His most recent work, including his dissertation, is centered on developing formal retrieval frameworks and

applying statistical language models to text retrieval, especially in directions such as personalized search and semi-structured information retrieval. He has served on the program committee for ACM SIGIR 2003, ACM SIGIR 2004, ACL 2003, ACM CIKM 2003. He is the IR program co-chair for ACM CIKM 2004. He is a recipient of the 2004 NSF CAREER award.