Book Review

Statistical Language Models for Information Retrieval

ChengXiang Zhai
University of Illinois at Urbana Champaign

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst), volume 1, 2008; xiii+125 pp, Princeton, NJ; paperbound, ISBN 978-1-59829-590-0, $40.00; ebook, ISBN 978-1-59829-591-7, $30.00 or by subscription

Reviewed by
Eric Gaussier
University Joseph Fourier & LIG

The past decade has seen a steady growth of interest in statistical language models for information retrieval, and much research work has been conducted on this subject. This book by ChengXiang Zhai summarizes most of this research. It opens with an introduction covering the basic concepts of information retrieval and statistical language models, presenting the intuitions behind these concepts. This introduction is then followed by a chapter providing an overview of:

- classical information retrieval models, from similarity-based models to probabilistic relevance and inference models, and
- two retrieval frameworks, the axiomatic and decision-theoretic retrieval frameworks, both co-developed by the author.

This overview covers the main aspects of these models and frameworks, and allows the author to introduce notions that help position the statistical language model to be presented in the following chapters.

The remainder of the book is then devoted to the presentation of the statistical language models used in information retrieval, and their application to special tasks. Chapter 3 presents the standard, simple query likelihood retrieval model. After reviewing the basic idea behind this model (at the core of current statistical language models for information retrieval), the author presents the different event models that have been considered: Multinomial, Multiple Bernoulli, and Multiple Poisson. He then explains the strategy for parameter estimation and the different smoothing techniques one can rely on. This presentation is followed by a discussion of the relation between smoothing and *tf-idf* weighting, which paves the way for the two-stage smoothing method presented in the following section. This chapter is very well written and presents, in a clear yet complete way, the fundamentals of the query likelihood retrieval model.

The following chapter, entitled Complex Query Likelihood Retrieval Model, is devoted to extensions of the simple query likelihood model of Chapter 3. In particular, the author reviews document-specific smoothing methods, based on document clustering and document expansion, the use of n-gram models and Markov random fields, as well as the full Bayesian query likelihood and the translation models. The intent here is not to provide a detailed description of these elements, but rather to give an overview, and pointers to extensions and models related to the query likelihood retrieval model.

A major conceptual drawback of the query likelihood retrieval model lies in the fact that feedback cannot be naturally accommodated. Indeed, in this model, a query is seen as a sample from the document model; adding terms according to a completely different process renders the "sample view" inadequate. In order to better accommodate feedback, the Kullback–Leibler divergence retrieval model has been introduced. In this model, a query model (associated with a word probability distribution) and a document model (also associated with a word probability distribution) are compared with the Kullback–Leibler divergence, a form of probabilistic distance. Chapter 5, entitled Probabilistic Distance Retrieval Model, is devoted to the presentation of this model, with a complete description of how to estimate query models and how to account for feedback (positive and negative). This chapter is slightly more technical than the previous ones. It, however, provides a very accurate and complete description of the Kullback–Leibler model—currently the most widely used statistical language model in information retrieval.

The next two chapters (Chapters 6 and 7) cover the application of these models to special aspects of the retrieval process. In Chapter 6, the author briefly reviews several retrieval tasks: cross-lingual information retrieval, distributed information retrieval, structured document retrieval, personalized and context-sensitive search, expert finding, passage retrieval, and sub-topic retrieval. Although some tasks are reasonably well covered (e.g., cross-lingual information retrieval), others are less detailed (e.g., distributed information retrieval), the author only providing a brief synthesis of the research conducted about the use of language models. The last chapter of the book, Chapter 7, entitled Language Models for Latent Topic Analysis, presents two widely used topic models, namely PLSA (Probabilistic Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation), and their application to information retrieval. The presentation of both PLSA and LDA is clear and precise. A section devoted to the extensions of these models furthermore provides a lot of pointers to related studies. Although topic models fit within the general definition adopted here for language models (*probability distribution over word sequences*, page 6), their goal radically differs from that of the statistical models presented in the previous chapters, so that choosing to devote a chapter to topic models may seem surprising at first. The importance of topic models and their use in various places of the retrieval process, however, justifies this choice. A conclusion summarizing the position of statistical language models with respect to traditional information retrieval models, and synthesizing the research progress on statistical language models over the last decade, closes the book.

Because of the focus of the book on theoretical developments and models, it definitely targets researchers and Ph.D. students in information retrieval who already possess some understanding of probability theory and statistical language models. Teachers may use it for a course on information retrieval, but only as a complement to general information retrieval books. Although most statistical approaches to information retrieval are mentioned in this book, one may regret the relatively poor treatment of the Divergence from Randomness (DFR) approach (Amati and Van Rijsbergen 2002). This approach has been shown to yield state-of-the-art results on several collections and could have been covered in more detail in Chapter 2 (Overview of Information Retrieval Models), all the more so since a formal relation between statistical language and DFR models was recently established (Clinchart and Gaussier 2009).

In conclusion, this book presents statistical language models for information retrieval in a thorough and clear manner. All the aspects of such models are very well presented and all the subtleties fully addressed. Furthermore, each chapter contains many pointers to related research, as well as a summary synthesizing the main results

presented. The part on the application to special tasks (Chapters 6 and 7) is not covered as deeply as the other parts, but does provide useful information and pointers. All in all, this is a very interesting book, both clear and complete, and written by a researcher who has been an important contributor to the field. Indeed, a substantial part of the material presented in the book was in fact developed or co-developed by the author himself.

This book review was edited by Pierre Isabelle.

References

Amati, Gianni and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389.

Clinchant, Stéphane and Eric Gaussier. 2009. Bridging language modeling and divergence from randomness models: A log-logistic model for IR. In *Advances in Information Retrieval: Proceedings of the 2nd International Conference on the Theory of Information Retrieval*. Cambridge, UK; Lecture Notes in Computer Science number 5766, Springer, pages 54–65.

Eric Gaussier is Professor at the University J. Fourier, Grenoble, France. His research focuses on statistical learning and modeling for textual information access. Eric Gaussier's address is: LIG, Bâtiment B - 385, Avenue de la Bibliothèque - BP 53 - 38041, Grenoble Cedex 9, France; e-mail: eric.gaussier@imag.fr.