Preface

The language modeling approach to information retrieval (IR) is a new framework that has been proposed and developed within the past five years, although its roots in the IR literature go back more than twenty years. Research carried out at a number of sites has confirmed that the language modeling approach is a theoretically attractive and potentially very effective probabilistic framework for building IR systems.

The central computational device in this framework is a *language model* - a probabilistic model for generating natural language text. The most familiar and basic language models are simply "unigram" word models, built in terms of the relative frequencies of the words appearing in a document. More sophisticated language models account for word order, phrases, and the change in language statistics in time and across document collections.

The use of language models is attractive for several reasons. For example, building an IR system using language models allows us to reason about the design and empirical performance of the system in a principled way, using the tools of probability theory. In addition, we can leverage the tremendous amount of work that has been carried out in the speech recognition community in the past thirty years on such issues as smoothing and combining language models for multiple topics and collections. The language modeling approach applies naturally to a wide range of information system technologies, such as *ad hoc* and distributed retrieval, cross-language IR, summarization and filtering, and, possibly, question answering. Language models can potentially be used to provide an integrated representation framework across documents, topics, collections, languages, queries, and users.

This workshop has two goals. The first is to promote the exchange of ideas among researchers using language modeling and other probabilistic models for IR research and development projects. The second is to gather feedback on the design of a language modeling toolkit for IR and related research. The development of this toolkit would encourage more research groups to contribute to this area and should lead to more rapid development of the related technologies.

The workshop has 32 participants from 5 countries. The 20 presentations cover a broad range of topics within the general area of probabilistic and language models. To further promote this area, we intend to produce a book containing a collection of papers after the workshop.

Enjoy the meeting and your stay in Pittsburgh.

Bruce Croft
Jamie Callan
John Lafferty
Workshop Organizers