Workshop on Language Modeling and Information Retrieval

May 31-June 1 2001
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

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.

Two groups working in this area at the University of Massachusetts and Carnegie-Mellon University, with the sponsorship of the Advanced Research and Development Activity in Information Technology (ARDA), recently organized a workshop in this area. The workshop had two goals. The first was to promote the exchange of ideas among researchers using language modeling and other probabilistic models for IR research and development projects. The second was 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 had 32 participants from 5 countries. The 20 presentations covered a broad range of topics within the general area of probabilistic and language models. The papers are available online at `http://la.lti.cs.cmu.edu/callan/Workshops/lmir01/`.

Some of the issues that received particular attention at the workshop included the role of relevance in language modeling approaches to IR and their relationship to other probabilistic retrieval models, the differences between a “query-likelihood” and a “document-likelihood” view of retrieval, the relative performance of smoothing techniques, the benefits of smoothing using document models compared to corpus-based topic models (local vs. global smoothing), techniques for incorporating more information beyond simple unigram probabilities in the language models and how these new features may be useful for retrieval, and applications of language modeling to summarization and categorization.

Robertson and Sparck Jones presented papers that challenged the language modeling (LM) community to show how their approaches are related to the probability ranking principle and retrieval models that estimate probability of relevance. They also mentioned the difficulty of describing important processes such as relevance feedback in current LM approaches. The papers by Lafferty and Zhai, and Lavrenko addressed these issues directly and suggested new forms of the LM approach to retrieval that were more closely related to previous probabilistic retrieval models. Both of these approaches moved away from estimating the probability of generating query text (the query-likelihood model) to estimating the probability of generating document text (document-likelihood) or comparing query and document language models directly. Both papers also presented techniques for estimating query models from very sparse data. Fuhr showed how the language modeling approach can be related to other probabilistic retrieval models in the framework of uncertain inference. Despite the contributions of these and other papers, the role of relevance continued to be a major theme throughout the workshop.

A number of people discussed the importance of estimation techniques. Much discussion centered on how this is the central issue of the LM approach and the similarities and differences to tf.idf weights. LM approaches have consistently outperformed systems using tf.idf weights even with relatively simple estimation techniques, but the evidence is growing that substantially better results may be possible. Greiff and Ponte discussed the importance of variance reduction. Zhai and Lafferty presented the results of a large number of retrieval experiments using different smoothing techniques. These experiments demonstrated the large range of performance that is possible depending on the estimation technique used. Hoffman discussed how latent classes or topic models can be incorporated into the language modeling framework. These topic models provide a form of smoothing based on reducing the dimensionality of the corpus. Peters discussed a similar approach based on clustering. Zhu and Rosenfeld showed how an exponential model approach combined with discriminative techniques can be used to extend the features used in the language models for retrieval beyond bags of words. The type of features that are need for effective retrieval in a language modeling framework was another of the major themes in the workshop discussions.

In the applications area, Spitters and Kraaij, and Allan, Gupta and Khandelwal discussed how language modeling approaches could be used for topic detection and tracking. Allan focused on the interesting task of detecting novelty in news stories on a given topic. Schwartz described a language modeling framework for selecting topic words for

document summaries. Larson described an approach to modeling sub-words for spoken document retrieval. Teahan and Harper generated a considerable amount of discussion with their results on categorization using a compression-based language model that captures statistics about character n-grams rather than words.

Finally, Manmatha and Zhang described similar approaches to modeling the probabilistic processes underlying retrieval and filtering scores and using those models to improve system performance.

Although this is a very “dry” summary, the workshop was one of the liveliest that many of us had attended for some time. It is clear that a lot of interest is being generated in this area, and that there are many issues that will not be resolved easily. The experimental results presented in many of the talks also verified that the interest is justified in terms of potential benefits to effectiveness for retrieval and other tasks.

A first version of a language modeling toolkit was also described and discussed at the workshop. This toolkit is scheduled to be released for general use around the time of SIGIR 2001 in Finland. A second workshop is planned for a similar date in the Boston area next year.

Bruce Croft
Jamie Callan
John Lafferty
Workshop Organizers