Variations on Language Modeling for Information Retrieval

Wessel Kraaij
TNO TPD
kraaijw@acm.org

Search engine technology builds on theoretical and empirical research results in the area of information retrieval (IR). This dissertation makes a contribution to the field of language modeling (LM) for IR, which views both queries and documents as instances of a unigram language model and defines the matching function between a query and each document as the probability that the query terms are generated by the document language model. The work described is concerned with three research issues.

The first research question addressed is how linguistic resources can be optimally combined with statistical language models. A case study on embedding morphological normalization for Dutch shows that complex models for matching in word form space are less effective than a simple model based on matching in the reduced feature space of word stems. A case study on cross-language information retrieval (CLIR) shows that probabilistic retrieval models with fully integrated statistical translation perform significantly better than the frequently applied synonym-based approach. A crucial element is the fact that the probabilistic models can accommodate multiple weighted translation variants, which is especially effective when translations are derived from parallel corpora.

The second research issue is an investigation of the hypothesis that it should be possible to formulate a single LM-based document ranking formula, which is effective for both topic tracking and ad hoc search. The first task differs from the latter by the fact that ranking score distributions for different topics must be comparable on an absolute scale. A variant model which meets this criterion is proposed and its relationship to the classical odds-of-relevance model and the Kullback-Leibler divergence is explained. The model is based on the reduction in cross-entropy associated with a certain document model in comparison to a background model. Besides being an adequate and unifying model for topic tracking and ad hoc search, the cross-entropy based approach also allows for intuitive modeling of the CLIR task by mapping either the query or document language model onto a language model in a different language.

The final research issue concerns a more general problem for IR researchers, namely statistical validation of experimental results. Standard statistical significance tests are reviewed and their underlying assumptions about the properties of test data are validated on actual data from IR experiments. A set of guidelines is presented, which contribute to an improved methodological framework. In addition, a fairly comprehensive discussion of state-of-the art IR models and techniques is included, which can be read as a tutorial text.

ISBN 90-75296-09-6; Taaluitgeverij Neslia Paniculata, Enschede
ISSN 1381-3617; No. 04-62 (CTIT Ph.D. -thesis series No. 04-62)
PDF version available at `http://vcard.acm.org/~kraaijw`