November 15, 2020

Language Modeling

A historical review

The first language modeling approach is published in 1998 by Jay M. Ponte and W. Bruce Croft. They proposed an approach to retrieval based on probabilistic language modeling, which is non-parametric and integrates document indexing and document retrieval into a single model. At that time, many of the indexing models make assumptions about the data: one is the parametric assumption, the other is that documents are members of pre-defined classes. They argued that they were both unwarranted.

The phrase 'language model' was first used by speech recognition community to refer to a probability distribution that captures the statistical regularities of the generation of language. For the purpose of document retrieval, one can model occurrences at the document level without regard to sequential effects and will be the approach taken by Ponte and Croft. They infer a language model for each document and to estimate the probability of generating the query according to each of these models. Then rank the documents according to these probabilities. In this approach, collection statistics such as term frequency, document length and document frequency are integral parts of the language model and are not heuristically as in many other approaches.

Since the pioneering work by Ponte and Croft, a lot of progress has been made in studying the language modeling approaches to IR, which we briefly reviewed here. The following is an list of some most important developments:

Framework and justification for using LMs for IR have been established: The query likelihood retrieval method has been shown to be a well-justified model according to the probability ranking principle. General frameworks such as the risk minimization framework and the generative relevance framework offer road maps for systematically applying language models to retrieval problems.

Many effective models have been developed and they often work well for multiple tasks:

The KL-divergence retrieval model, which covers the query likelihood retrieval model, has been found to be a solid and empirically effective retrieval model that can easily incorporate many advanced language models; many methods have been developed to improve estimation of query language models. Dirichlet prior smoothing has been recognized as an effective smoothing method for retrieval. The KL-divergence retrieval model combined with Dirichlet prior smoothing represents the state-of-the-art baseline method for the language modeling approaches to IR. The translation model proposed in enables handling polysemy and synonyms in a principled way with a great potential for supporting semantic information retrieval. The relevance model offers an elegant solution to the estimation problem in the classical probabilistic retrieval model as well as serves as an effective feedback method for the KL-divergence retrieval model. Mixture unigram language models have been shown to be very powerful and can be useful for many purposes such as pseudo feedback, improving model discriminativeness, and modeling redundancy.

It has been shown that completely automatic tuning of parameters is possible for both non-feedback retrieval and pseudo feedback. LMs can be applied to virtually any retrieval task with great potential for modeling complex IR problems.

It has been a long-standing challenge in IR research to develop robust and effective retrieval models. As a new generation of probabilistic retrieval models, language modeling approaches have several advantages over traditional retrieval models such as vector-space model and the classical probabilistic retrieval model:

First, these language models generally have a good statistical foundation. This makes it possible to leverage many established statistical estimation methods to set parameters in a retrieval function. Following rigorous statistical modeling also forces any assumptions to be made explicit. A good understanding of such assumptions often helps diagnose the weakness and strength of a model and thus better interpret experiment results. Second, they provide a principled way to address the critical issue of the text representation and term weighting. The issue of term weighting has long been recognized as critical, but before language modeling approaches were proposed, this issue had been traditionally addressed mostly in a heuristic way. Language models, multinomial unigram language models particular, can incorporate term frequencies and document length normalization naturally into a probabilistic model. While such connection has also been made in the classic probabilistic retrieval model, the estimation of parameters was not addressed as seriously as in the language models. Third, language models can often be more easily adapted to model various kinds of complex and special retrieval problems than traditional models. The Benefit has largely come from the availability of many well-understood statistical models such as finite mixture models, which can often be estimated easily by using the EM algorithm.

However, the language modeling approaches also have some deficiencies as compared with traditional models: First, there is a lack of explicit discrimination in most of the language models developed so far. Such a lack of discrimination is indeed a general problem with all generative models as they are designed to describe what the data looks like rather than how the data differs. Second, the language models have been found to be less robust than the traditional TF-IDF model in some cases and can perform poorly or be very sensitive to parameter setting. Third, some sophisticated language models can be computationally expensive, which may limit their uses in large-scale retrieval applications.