Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 80 (2016) 49–59

www.elsevier.com/locate/specom

Exploring the use of unsupervised query modeling techniques for speech recognition and summarization

Kuan-Yu Chen[a], Shih-Hung Liu[a], Berlin Chen[b,*], Hsin-Min Wang[a], Hsin-Hsi Chen[c]

[a] *Institute of Information Science, Academia Sinica, Taipei, Taiwan*
[b] *Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan*
[c] *Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan*

Received 9 November 2015; received in revised form 9 March 2016; accepted 13 March 2016
Available online 26 April 2016

Abstract

Statistical language modeling (LM) that intends to quantify the acceptability of a given piece of text has long been an interesting yet challenging research area. In particular, language modeling for information retrieval (IR) has enjoyed remarkable empirical success; one emerging stream of the LM approach for IR is to employ the pseudo-relevance feedback process to enhance the representation of an input query so as to improve retrieval effectiveness. This paper presents a continuation of such a general line of research and the major contributions are three-fold. First, we propose a principled framework which can unify the relationships among several widely-cited query modeling formulations. Second, on top of this successfully developed framework, two extensions have been proposed. On one hand, we present an extended query modeling formulation by incorporating critical query-specific information cues to guide the model estimation. On the other hand, a word-based relevance modeling has also been leveraged to overcome the obstacle of time-consuming model estimation when the framework is being utilized for practical applications. In addition, we further adopt and formalize such a framework to the speech recognition and summarization tasks. A series of experiments reveal the empirical potential of such an LM framework and the performance merits of the deduced models on these two tasks.

Keywords: Query modeling; Language modeling; Information Retrieval; Speech recognition; Summarization.

1. Introduction

Following the exponentially growing popularity of the Internet and the ubiquity of social web communications, tremendous volumes of multimedia content, such as broadcast radio and television programs, digital libraries and so on, are made available to the public. This has spurred a surge of research on speech-based multimedia content understanding and organization over the past two decades. By virtue of the developed processing techniques, a variety of functionalities were created to help distill important content from multimedia collections, or provide locations of important speech segments in a video along with their corresponding transcripts, for users to listen to or to digest.

On a separate front, statistical language modeling (LM) (Jelinek, 1999; Jurafsky and Martin, 2008; Zhai, 2008), which manages to quantify the acceptability of a given word sequence in natural language, or capture the statistical characteristics of a given piece of text, has been proved to offer both efficient and effective modeling abilities in many practical applications of natural language processing and speech recognition (Ponte and Croft, 1998; Jelinek, 1999; Huang et al., 2001; Jurafsky and Martin, 2008; Furui et al., 2012; Liu and Hakkani-Tur, 2011). In particular, the LM-based modeling paradigm was first introduced for information retrieval (IR) problems in the late 1990s, indicating very good potential, and was subsequently extended in a wide array of follow-up studies. One typical realization of this paradigm for IR is to access the degree of relevance between a query and a document by computing the likelihood that the query is generated by the document (usually referred to as the query-likelihood

* Corresponding author. Tel.: +886 2 7734 6672; fax: +886 2 29322378.
E-mail address: berlin@csie.ntnu.edu.tw (B. Chen).

http://dx.doi.org/10.1016/j.specom.2016.03.006
0167-6393/

approach) (Zhai, 2008; Baeza-Yates and Ribeiro-Neto, 2011). A document is deemed to be relevant to a given query if the corresponding document model is more likely to generate the query. Alternatively, the Kullback–Leibler divergence measure (denoted by KLM for short hereafter), which quantifies the degree of relevance between a document and a query from a more rigorous information-theoretic perspective, has been proposed (Lafferty and Zhai, 2001; Zhai and Lafferty, 2001a; Baeza-Yates and Ribeiro-Neto, 2011). KLM not only can be thought as a natural generalization of the query-likelihood approach, but also has the additional advantage of being able to accommodate extra information cues to improve the performance of document ranking. For example, a main challenge facing the LM-based modeling paradigm is that since a given query usually consists of few words, the true information need is hard to be inferred from the surface statistics of a query. As such, one emerging stream of thought for KLM is to employ the pseudo-relevance feedback process to construct an enhanced query model (or representation) so as to achieve better retrieval effectiveness (Lavrenko and Croft, 2001; Zhai and Lafferty, 2001; Tao and Zhai, 2006; Hiemstra et al., 2004; Lv and Zhai, 2009; Carpineto and Romano, 2012; Lee and Croft, 2013; Clinchant and Gaussier, 2013; Chen et al., 2013b, 2014).

This paper presents a continuation of this general line of research and its main contributions are highlighted as follows: (1) we analyze several widely-used query models and then propose a principled framework to unify the relationships among them; (2) on top of the successfully developed query models, we propose two extended modeling formulations by incorporating additional query-specific information cues to guide the model estimation, or by introducing word-word relevance cues to mitigate the computation time problem in realistic applications; (3) we explore a novel use of these query models by adapting them to the speech recognition and summarization tasks. As we will see, a series of experiments indeed confirm the effectiveness of the proposed models on these two disparate tasks.

2. Language modeling framework for IR

2.1. Query likelihood measure (QLM)

Language modeling (LM) has emerged as a promising paradigm for building information retrieval systems (Chen, 2009; Chia et al., 2010, 2012). This is due to the fact that the LM-based paradigm has inherently neat probabilistic foundation and excellent performance (Zhai, 2008). The fundamental formulation of such a paradigm for information retrieval is to compute the conditional probability $P(Q|D)$, i.e., the likelihood that a query Q is generated by each document D (the so-called query likelihood measure). A document D is deemed to be relevant with respect to the query Q if the corresponding document model is more likely to generate the query. If the query Q is treated as a sequence of words, $Q = w_1, w_2, \ldots, w_L$, where the query words are assumed to be conditionally independent given the document D and their order is also assumed to be of no importance (i.e., the so-called "*bag-of-words*" assumption), the likelihood measure $P(Q|D)$ can be further decomposed as a product of the probabilities of the query words generated by the document (Zhai, 2008):

$$P(Q|D) = \prod_{l=1}^{L} P(w_l|D)\,, \tag{1}$$

where $P(w_l|D)$ is the likelihood of generating w_l by the document D (also known as the document model). The simplest way to construct $P(w_l|D)$ is based on literal term matching (Lee and Chen, 2005), through using the unigram language model (ULM). To this end, each document D can, respectively, offer a unigram distribution for observing any given word w, which is parameterized on the basis of the empirical counts of words occurring in the document with the maximum likelihood (ML) estimator (Jelinek, 1999; Zhai, 2008):

$$P(w|D) = \frac{c(w, D)}{|D|}, \tag{2}$$

where $c(w,D)$ denotes the number of times that word w occurs in the document D and $|D|$ is used to designate the number of words in the document. The document model is further smoothed by a background unigram language model estimated from a large general collection to model the general properties of the language as well as to avoid the problem of zero probability (Zhai, 2008). However, how to strike the balance between these two probability distributions is actually a matter of judgment, or trial and error.

2.2. Kullback–Leibler divergence measure (KLM)

Another effective realization of the LM-based paradigm for IR is the Kullback–Leibler divergence measure (KLM), which determines the degree of relevance between a document and a query from a rigorous information-theoretic perspective. Two different language models are involved in KLM: one for the document and the other for the query. KLM assumes that words in the query are random draws from a language distribution that describes the information need of a user, and words in the relevant documents should also be drawn from the same distribution. The divergence of the document model with respect to the query model is defined by

$$\mathrm{KL}(Q||D) = \sum\nolimits_{w\in V} P(w|Q) \log \frac{P(w|Q)}{P(w|D)}. \tag{3}$$

In addition, it is easy to show that the KL-divergence measure will give the same ranking as the ULM model (cf. Eq. (1)) when the query language model is simply derived with the ML estimator (Chen et al., 2012):

$$\begin{aligned} -\mathrm{KL}(Q\|D) &\overset{\text{rank}}{=} \sum\nolimits_{w\in V} P(w|Q) \log P(w|D) \\ &= \sum\nolimits_{w\in V} \frac{c(w, Q)}{|Q|} \log P(w|D) \\ &\overset{\text{rank}}{=} \sum\nolimits_{w\in V} c(w, Q) \log P(w|D) \\ &= \log P(Q|D) \\ &\overset{\text{rank}}{=} P(Q|D). \end{aligned} \tag{4}$$