Statistical Language Modeling For Information Retrieval

Xiaoyong Liu and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
{xliu, croft}@cs.umass.edu

1. Introduction

This chapter reviews research and applications in statistical language modeling for information retrieval (IR) that has emerged within the past several years as a new probabilistic framework for describing information retrieval processes. Generally speaking, statistical language modeling, or more simply, language modeling (LM), refers to the task of estimating a probability distribution that captures statistical regularities of natural language use. Applied to information retrieval, language modeling refers to the problem of estimating the likelihood that a query and a document could have been generated by the same language model, given the language model of the document and with or without a language model of the query.

The root of statistical language modeling dates back to the beginning of the 20th century when Markov tried to model letter sequences in works of Russian literature (Manning and Schütze, 1999). Zipf (1929, 1932, 1949, 1965) studied statistical properties of text and discovered that the frequency of words decays as a power function of its rank. However, it was Shannon's work (Shannon, 1951) that inspired later research in this area. In 1951, eager to explore the applications of his newly founded information theory to human language, Shannon used a prediction game that involved n-grams to investigate the information content of English text. He evaluated n-gram models' performance by comparing their cross-entropy on text with the true entropy estimated using predictions made by human subjects. For many years, statistical

language models have been used primarily for automatic speech recognition. Since 1980 when the first significant language model was proposed (Rosenfeld, 2000), statistical language modeling has become a fundamental component of speech recognition, machine translation, spelling correction, and so forth. It has also proven useful for natural language processing tasks such as natural language generation and summarization. In 1998, it was introduced to information retrieval and has opened up new ways of thinking about the retrieval process.

The first uses of language modeling approach for IR focused on its empirical effectiveness using simple models. In the basic approach, a query is considered generated from an "ideal" document that satisfies the information need. The system's job is then to estimate the likelihood of each document in the collection being the ideal document and rank them accordingly. This query-likelihood retrieval model, first proposed by Ponte & Croft (1998), and later described in terms of a "noisy channel" model by Berger & Lafferty (1999), has produced results that are at least comparable to the best retrieval techniques previously available. The basic model has been extended in a variety of ways. For example, documents have been modeled as mixtures of topics (Hofmann, 1999a) and phrases are considered (Song & Croft, 1999). Progress has also been made in understanding the formal underpinnings of the statistical language modeling approach, and comparing it to traditional probabilistic approaches. Connections were found and differences identified. Recent work has seen more sophisticated models developed that are more closely related to the traditional approaches. For example, a language model that explicitly models relevance (Lavrenko & Croft, 2001) has been proposed, and a risk-minimization framework based on Bayesian decision theory has been developed (Lafferty & Zhai, 2001a). Successful applications of the LM approach to a number of retrieval tasks have also been reported, including cross-lingual retrieval (Xu, et al., 2001; Lavrenko et al., 2002) and

distributed retrieval (Xu & Croft, 1999; Si et al., 2002). Research carried out by a number of groups has confirmed that the language modeling approach is a theoretically attractive and potentially very effective probabilistic framework for studying information retrieval problems (Croft & Lafferty, 2003). This empirical success and the overall potential of the approach have also triggered the *LEMUR*[1] project.

As a new family of probabilistic retrieval models, language models for IR share the theoretical foundations underlying the general probabilistic IR work. Numerous authors have contributed to the theoretical discussions of probabilistic retrieval including Maron & Kuhns (1960), Cooper (1978, 1995), Robertson & Sparck Jones (1976), Robertson (1977), van Rijsbergen (1979, 1992), and Sparck Jones et al. (2000a, 2000b), to name a few. The well known probabilistic IR models include the Robertson & Sparck Jones model (1976), the Croft & Harper model (1979), the Fuhr model (1989), and the inference network model (Turtle & Croft, 1991). Detailed treatment of these earlier probabilistic IR theories and approaches is beyond the scope of this paper. Excellent renderings of this topic can be found in (van Rijsbergen, 1979) and (Baeza-Yates & Ribeiro-Neto, 1999). Readers are encouraged to consult them for more information.

This review brings together contemporary research on statistical language modeling and smoothing for retrieval of written text. The review does not cover language-modeling techniques that are developed for speech recognition and other language technologies but have not been applied to text retrieval. It also limits the discussion on earlier probabilistic IR work for which there is a wealth of literature. This is the first ARIST review of statistical language modeling for

[1] This is a collaborative project of the University of Massachusetts and Carnegie-Mellon University, with the sponsorship of the Advanced Research and Development Activity in Information Technology (ARDA), in developing an open-source toolkit for language modeling in information retrieval. The toolkit is available at http://www.cs.cmu.edu/~lemur/.

IR, but Rosenfeld (2000) reviewed language modeling techniques for speech recognition and other domains, and Chen & Goodman (1998) provided a survey of smoothing techniques developed for those domains.

The rest of the chapter is organized as follows. Section 2 introduces statistical language modeling in more detail and overviews major LM techniques for IR. Section 3 discusses various smoothing strategies used in language models for IR. Section 4 draws comparisons between LM and traditional probabilistic IR approaches. Applications of LM to various retrieval tasks are discussed in section 5. The review concludes in section 6 with some observations regarding future research directions.

2. Language models for IR

A statistical language model is a probability[2] distribution over all possible sentences or other linguistic units in a language (Rosenfeld, 2000). It can also be viewed as a statistical model for generating text. The task of language modeling, in general, answers the question: how likely the *i*th word in a sequence would occur given the identities of the preceding *i*-1 words? In most applications of language modeling, such as speech recognition and information retrieval, the probability of a sentence is decomposed into a product of *n-gram* probabilities.

Let's assume that S denotes a specified sequence of k words,

$$S = w_1, w_2, ..., w_k$$

An n-gram language model considers the word sequence S to be a Markov process with probability

[2] The formulation of language models have been based on probability theory. There are a number of theories of what probability means, and the differences can have an effect on how probabilistic models are interpreted. A good discussion of the various theories can be found in (Good, 1950).

$$P_n(S) = \prod_{i=1}^{k} P(w_i \mid w_{i-1}, w_{i-2}, w_{i-3}, ..., w_{i-n+1})$$

where n refers to the order of the Markov process. When $n = 2$ we call it a bigram language model which is estimated using information about the co-occurrence of pairs of words. In the case of n=1, we call it a unigram language model which uses only estimates of the probabilities of individual words. For applications such as speech recognition or machine translation, word order is important and higher-order (usually trigram) models are used. In information retrieval, the role of word order is less clear and unigram models have been used extensively.

To establish the word n-gram language model, probability estimates are typically derived from frequencies of n-gram patterns in the training data. It is common that many possible word n-gram patterns would not appear in the actual data used for estimation, even if the size of the data is huge and the value of n is small. As a consequence, for rare or unseen events the likelihood estimates that are directly based on counts become problematic. This is often referred to as the data sparseness problem. Smoothing is used to address this problem and has been an important part in any language model. We will save it for a detailed discussion in the next section. In this section, we focus our discussion of various language models on their conceptual similarities and differences.

Evaluation of language models in other domains has typically been done using a measure called "perplexity" (Manning and Schütze, 1999). This measure is directly related to entropy. Entropy measures the average uncertainty present for a random variable (Cover & Thomas, 1991). The more knowledge or structure a model captures, the lower the uncertainty, or entropy will be. Models with lower entropy can therefore be considered better. In *ad hoc* IR, performance of retrieval models has mostly been evaluated based on precision and recall. The average precision measure combines precision and recall into a single-number summary. Baeza-Yates &

Ribeiro-Neto (1999) give a good discussion on these measures and their appropriateness. In order for the performance of language models to be directly comparable to that of other retrieval models, researchers have taken the average precision measure as the method of choice for evaluation. Throughout this paper when we discuss retrieval performance we refer to that measured in average precision. We now begin our discussion on models.

Query likelihood model. The basic approach for using language models for IR assumes that the user has a reasonable idea of the terms that are likely to appear in the "ideal" document that can satisfy his/her information need, and that the query terms the user chooses can distinguish the "ideal" document from the rest of the collection (Ponte & Croft, 1998). The query is thus generated as the piece of text representative of the "ideal" document. The task of the system is then to estimate, for each of the documents in the collection, which is most likely to be the ideal document. That is, we calculate:

$$\arg\max_{D} P(D \mid Q) = \arg\max_{D} P(Q \mid D)P(D) \qquad (1)$$

where Q is a query and D is a document. The prior probability $P(D)$ is usually assumed to be uniform and a language model $P(Q|D)$ is estimated for every document. In other words, we estimate a probability distribution over words for each document and calculate the probability that the query is a sample from that distribution. Documents are ranked according to this probability. This is generally referred to as the *query-likelihood* retrieval model and was first proposed by Ponte & Croft (1998). In their paper, Ponte & Croft take a *multi-variate Bernoulli* approach to approximate $P(Q|D)$. They represent a query as a vector of binary attributes, one for each unique term in the vocabulary, indicating the presence or absence of terms in the query. The number of times that each term occurs in the query is not captured. There are a couple of assumptions behind this approach: 1) the *binary* assumption: all attributes are binary. If a term

occurs in the query, the attribute representing the term takes the value of 1. Otherwise, it takes the value of 0. And, 2) the *independence* assumption: terms occur independently of one another in a document. These assumptions are the same as those underlie the *binary independence* model proposed in earlier probabilistic IR work (Robertson & Sparck Jones, 1976; van Rijsbergen, 1977). Based on these assumptions, the query likelihood *P(Q|D)* is thus formulated as the product of two probabilities – the probability of producing the query terms and the probability of not producing other terms.

$$P(Q \mid D) = \prod_{w \in Q} P(w \mid D) \prod_{w \notin Q} (1.0 - P(w \mid D)) \qquad (2)$$

where *P(w|D)* is calculated by a non-parametric method that makes use of the average probability of *w* in documents containing it and a risk factor. For non-occurring terms, the global probability of *w* in the collection is used instead. It is worth mentioning that collection statistics such as term frequency and document frequency are integral parts of the language model and not used heuristically as in traditional probabilistic and other approaches. In addition, document length normalization does not have to be done in an *ad hoc* manner as it is implicit in the calculation of the probabilities. This approach to retrieval, although very simple, has demonstrated superior performance to traditional probabilistic retrieval using the Okapi-style *tf-idf* weighting (Robertson et al., 1995) on TREC[3] test collections. An 8.74% improvement in performance (measured in average precision) is reported in the paper. This finding is important because with few heuristics the simple language model can do at least as well as one of the most successful probabilistic retrieval models previously available with heuristic *tf-idf* weighting.

[3] TREC stands for the Text REtrieval Conference. Co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA), the conference supports research in the information retrieval community by providing infrastructure (such as realistic test collections and appropriate evaluation procedures) for large-scale evaluation of various (text) retrieval methodologies. It also serves as a forum for the exchange of research ideas and for the discussion of research methodology. More information can be found at: http://trec.nist.gov/

In contrast to Ponte & Croft's approach, Hiemstra (1998), Miller et al. (1999), and Song & Croft (1999) employ a *multinomial* view of the query generation process. They treat the query Q as a sequence of independent terms (i.e. $Q = q_1, \ldots, q_m$), taking into account possibly multiple occurrences of the same term. The "ordered sequence of terms assumption" behind this approach states that both queries and documents are defined by an ordered sequence of terms (Hiemstra, 1998). A query of length k is modeled by an ordered sequence of k random variables, one for each term occurrence in the query. While this assumption is not usually made in traditional probabilistic IR work, it has been essential for many statistical natural language processing tasks (e.g. speech recognition). Based on this assumption, the query probability can be obtained by multiplying the individual term probabilities.

$$P(Q \mid D) = \prod_{i=1}^{m} P(q_i \mid D) \qquad (3)$$

where q_i is the ith term in the query. While through different theoretical derivations, these models all arrived at a similar way of computing $P(w|D)$ (with w denoting any term) - combining a component estimated from the document and one from the collection by linear interpolation (we refer to this formulation as the probability-weighted model in the rest of this review).

$$P(w \mid D) = \lambda P_{document}(w \mid D) + (1 - \lambda) P_{collection}(w) \qquad (4)$$

where λ is a weighting parameter between 0 and 1. This can also be viewed as a combination of information from a local source, i.e. the document, and a global source, i.e. the collection. The differences between those models reside in how $P_{document}(w|D)$ and $P_{collection}(w)$ are estimated. Hiemstra (1998) relates his model to the well-known *tf-idf* formulation by approximating $P_{document}(w|D)$ using the maximum likelihood of term w appearing in the document D, which can be thought of as based on term frequency, and estimating $P_{collection}(w)$ using document frequency information. That is,

$$P_{document}(w \mid D) \cong P_{ml}(w \mid D) = \frac{c(w, D)}{\sum_{w' \in D} c(w', D)} \quad (5)$$

where *c(w, D)* is the number of times *w* occurs in *D* and $\sum_{w' \in D} c(w', D)$ is total number of tokens in *D*. And

$$P_{collection}(w) \cong \frac{df(w)}{\sum_{w' \in V} df(w')} \quad (6)$$

where *V* is the vocabulary and *df(w)* is the document frequency of term *w*, i.e. the number of documents in which term *w* appears. A pilot experiment on the Cranfield test collection shows that this model outperforms the traditional vector space model with *tf-idf* and cosine normalization.

Miller et al. (1999) use a two state hidden Markov model (HMM) with one state representing choosing a word directly from the document and the other state representing choosing a word from general English. $P_{document}(w|D)$ is the output distribution for the "document" state, which is estimated in the same way as in equation (5). To approximate the probability distribution of the "general English" state, the sample distribution of the entire document collection is used and $P_{collection}(w)$ is estimated by the maximum likelihood of term *w* occurring in the collection

$$P_{collection}(w) \cong P_{ml}(w \mid C) = \frac{c(w, C)}{\sum_{w' \in V} c(w', C)} \quad (7)$$

where *c(w, C)* is the number of times *w* occurs in the entire collection, *V* is the vocabulary, and $\sum_{w' \in V} c(w', C)$ is the total number of tokens in the collection. If we substitute equations (5) and (7) into equation (4), we get the following estimate

$$P(w \mid D) = \lambda \frac{c(w, D)}{\sum_{w' \in D} c(w', D)} + (1 - \lambda) \frac{c(w, C)}{\sum_{w' \in V} c(w', C)} \quad (8)$$

This is the basic formulation of the HMM model proposed by Miller et al. and often referred to as the simple language model which has been used as the baseline language model in several studies (Lavrenko & Croft, 2001; Liu & Croft, 2002; Jin et al., 2002). Retrieval experiments on TREC test collections show that the simple two-state system can do dramatically better than the *tf-idf* measure. This work also shows the possibilities that commonly used IR techniques such as relevance feedback as well as prior knowledge of the usefulness of documents can be incorporated into language models. For example, the authors modify their basic model (in equation (8)) to allow for: 1) incorporating automatic relevance feedback by re-estimation of P(w|D) with additional information from top ranked documents; 2) adding a third, document-dependent bigram state to the HMM; 3) weighting the importance of different query sections[4] and using the weights as part of the model; and 4) varying the document prior *P(D)* (as in equation (1)) according to features that are predictive of the usefulness of the document such as source, length, and average word-length. The refined system, whether using just one technique or all above four techniques, has had further performance gains, and using all techniques together is found to be superior to using any one of them.

Rather than using maximum likelihood estimates for computing probabilities in equation (4), Song & Croft (1999) propose to use Good-Turing estimates. Their basic model is the unigram model presented by equation (4) with Good-Turing estimates. In their second model, the unigram model is combined with a bigram model through linear interpolation. They compare the performance of the combined model with their basic model and the Ponte & Croft (1998) model,

[4] In Miller et al. (1999), the queries are taken from TREC (Text REtrieval Conference) topics. A TREC topic typically consists of three sections: Title, Description, and Narrative, with increasing details about a given subject.

and the performance of language models with INQUERY, a probabilistic retrieval system based on the inference net model (Turtle, 1990). Experiments on TREC data sets show that the results of language models are comparable to that of INQUERY. However, their basic and combined models have produced similar results, and the improvement over the Ponte & Croft model has only been marginal.

Statistical translation model. Taking a different angle, Berger and Lafferty (1999) view a query as a distillation or translation from a document. The query generation process is described in terms of a "noisy channel" model. To determine the relevance of a document to a query, their model estimates the probability that the query would have been generated as a translation of that document. Documents are then ranked according to these probabilities. More specifically, the mapping from a document term *w* to a query term *q* is achieved by estimating translation models *t(q|w)*. Using translation models, the retrieval model becomes

$$P(Q \mid D) = \prod_{i=1}^{m} \sum_{w} t(q_i \mid w) P(w \mid D)$$

A notable feature of this model is an inherent query expansion component and its capability of handling the issues of synonymy (multiple terms having similar meanings) and polysemy (the same term having multiple meanings). However, as the translation models are context independent, their ability to handle the ambiguity of word senses is only limited. While significant improvements over the baseline language model through the use of translation models is reported, this approach is not without its weaknesses: the need of a large collection of training data for estimating translation probabilities, and inefficiency for ranking documents.

Building upon the ideas of Berger & Lafferty (1999), Jin et al. (2002) propose to construct language models of document titles and determine the relevance a document to a query by estimating the likelihood that the query would have been the title for the document. The title

of a document is viewed as a translation from that document and the title language model is regarded as an approximate language model of the query. Jin et al. (2002) first estimate a translation model by using all the document-title pairs in a collection. The translation model is then used for mapping a regular document language model to a title language model. In the final step, the title language model estimated for each document is used to compute the query likelihood, and documents are ranked accordingly. It has been shown empirically that the title language model outperforms the simple language model (given in equation (8)) as well as the traditional Okapi method.

Risk minimization framework. Lafferty & Zhai (2001a, 2001b) and Zhai (2002) develop a risk minimization framework based on Bayesian decision theory. In this framework, queries and documents are modeled using statistical language models, user preferences are modeled through loss functions, and retrieval is cast as a risk minimization problem. The similarity between a document and a query is measured by the Kullback-Leibler (KL) divergence between the document model and the query model.

$$KL(Q \parallel D) = \sum_{w \in V} P(w \mid Q) \log \frac{P(w \mid Q)}{P(w \mid D)} \qquad (9)$$

One important advantage of this framework over previous approaches is its capability of modeling not only documents but also queries directly though statistical language models. This makes it possible to set retrieval parameters automatically and improve retrieval performance through utilization of statistical estimation methods, which is not typically done with traditional retrieval methods. This framework bears resemblance to the classical probabilistic retrieval models and can accommodate existing language models proposed by Ponte & Croft (1998) and others. Lafferty & Zhai (2001a) also introduce the idea of estimating expanded query language models for which they use a Markov chain method to help overcome the limitations of the

translation models used by Berger & Lafferty (1999). In the follow-up work, Zhai & Lafferty (2002) suggest using two-stage language models to explicitly capture different influences of the query and document collection on the optimal setting of retrieval parameters. In the first stage, a document language model is estimated independent of the query. In the second stage, query likelihood is computed according to a query language model, which is based on the estimated document language model from the first stage and a query background language model. This approach is similar to the original query likelihood approach by Ponte & Croft (1998) in that it involves both estimation of a document language model and computation of the query likelihood. The difference lies in whether the query likelihood is computed directly using the estimated document model (as is done in the original approach) or using a query model that is based on the estimated language model (as is done in the two-stage approach). A two-stage smoothing method is developed in this approach to set retrieval parameters completely automatically. Empirical evaluations indicate that the two-stage smoothing method consistently gives performance that is comparable with or better than the best obtainable by a single-stage smoothing method which is usually achieved by an exhaustive search through the whole parameter space.

Relevance model. Instead of attempting to model the query generation process, Lavrenko & Croft (2001) explicitly model relevance, and put forward a novel technique that estimates a relevance model from the query alone, with no training data. Conceptually, the relevance model is a description of an information need or, alternatively, a description of the topic area associated with the information need. It is assumed that, given a collection of documents and a user query Q, there exists an unknown relevance model R that assigns the probabilities $P(w|R)$ to the word occurrence in the relevant documents. The relevant documents are random samples from the distribution $P(w|R)$. Both the query and the documents are samples from R. The essence of their

model is to estimate *P(w|R)*. Let *P(w|R)* denotes the probability that a word sampled at random from a relevant document would be the word *w*. If we know what documents are relevant, estimation of these probabilities would be straightforward, but in a typical retrieval environment we are not given any examples of relevant documents. Lavrenko & Croft (2001) and Lavrenko et al. (2002) suggest a reasonable way to approximate *P(w|R)* by using a joint probability of observing the word *w* together with query words $q_1, \ldots, q_m$ ($Q = q_1, \ldots, q_m$):

$$P(w \mid R) \approx P(w \mid Q) = \frac{P(w, q_1 \ldots q_m)}{P(q_1 \ldots q_m)} = \frac{P(w, q_1 \ldots q_m)}{\sum_{v \in vocabulary} P(v, q_1 \ldots q_m)} \qquad (10)$$

Two methods of estimating the joint probability $P(w, q_1, \ldots, q_m)$ are described in (Lavrenko & Croft, 2001). Both methods assume the existence of a set *U* of underlying source distributions from which $w, q_1, \ldots, q_m$ could have been sampled. They differ in their independence assumptions. *Method 1* assumes that all query words and the words in relevant documents are sampled from the same distribution thus *w* and $q_1, \ldots, q_m$ are mutually independent once we pick a source distribution *M* from *U*. If we assume *U* to be the universe of our unigram language models, one for each document in the collection, then we get:

$$P(w, q_1 \ldots q_m) = \sum_{M \in U} P(M) P(w, q_1 \ldots q_m \mid M) = \sum_{M \in U} P(M) \left(P(w \mid M) \prod_{i=1}^{m} P(q_i \mid M) \right) \qquad (11)$$

where *P(M)* denotes some prior distribution over the set *U* which is usually taken to be uniform and *P(w|M)* specifies the probability of observing *w* if we samples a random word from *M*. *P(w|M)* is computed using equation (8). *Method 2* assumes that the query words $q_1, \ldots, q_m$ are independent of each other but are dependent on *w*. That is

$$P(w, q_1 \ldots q_m) = P(w) \prod_{i=1}^{m} P(q_i \mid w)$$

The conditional probability $P(q_i|w)$ can be estimated by calculating the expectation over the universe U of our unigram models:

$$P(q_i \mid w) = \sum_{M_i \in U} P(q_i \mid M_i) P(M_i \mid w)$$

Here again an assumption is made that q_i is independent of w once a source distribution M_i is picked. However, the difference from the assumption made in *Method 1* is that each q_i is now allowed to have a separate M_i. Although *Method 2* is shown to be less sensitive to the choice of the universe of distributions U and slightly more effective for retrieval, the relative simplicity and decomposability of *Method 1* has often made it the method of choice for estimation when the relevance model is used (Lavrenko et al., 2002; Liu & Croft, 2002; Cronen-Townsend et al., 2002). Lavrenko et al. (2002) employ the KL divergence between the relevance model and the document model to rank documents. Documents with smaller divergence from the relevance model are considered more relevant. The relevance model presents a natural incorporation of query expansion into the language model framework. Significantly improved retrieval performance has been reported by Lavrenko & Croft (2001) over a simple baseline language model similar to that of equations (3) and (8).

3. Smoothing strategies

Virtually all language models developed for IR to date use some form of an n-gram. To derive n-gram probabilities from corpora, one natural solution is to use maximum likelihood (ML) estimation, that is,

$$P_{ml}(w_n \mid w_1, w_2, ..., w_{n-1}) = \frac{c(w_1, w_2, ..., w_n)}{c(w_1, w_2, ..., w_{n-1})}$$

where $c(w_1, w_2, ..., w_n)$ is the raw counts of the occurrences of the word sequence $w_1, w_2, ..., w_n$ in the corpus. For unigram, the maximum likelihood of a term *w* is simply the number of the occurrences of *w* in the corpus divided by the total number of tokens in the corpus. The name maximum likelihood estimate comes from the fact that it does not waste any probability mass on unseen events but rather maximizes the probability of observed events based on the training data while subject to the normal stochastic constraints (Manning & Schütze, 1999). However, it is common that many possible word n-gram patterns would not appear in the actual data, even if the size of data is very large and the value of *n* is small. This causes the ML probability estimates of the unseen n-grams to be zero which is rather extreme – the fact that we have not seen an n-gram does not make it impossible to occur. Even worse is that these zero probabilities propagate. For example, the basic LM approach to IR attempts to find the document D that maximizes the probability *P(Q|D)* for a given query *Q*, and that probability is generally computed by multiplying the probabilities of individual query terms. If a document is missing one or more of the query terms, it will receive a probability of zero even though it may be highly relevant. This is generally known as the sparse data problem, and sometimes referred to as the zero-frequency problem (Witten & Bell, 1991). Smoothing is used to address this problem and has become an indispensable part of any language model. Smoothing is so called because the techniques developed for this purpose tend to make the distributions more uniform, by pushing low probabilities (including zero probabilities) upward while balancing out the adjustments by pushing high probabilities downward, so that the total probability mass still sums up to one. While the sparse data problem is the most prominent reason for smoothing in LM for IR, it is not the only reason. There are other subtler roles that smoothing plays including combining multiple knowledge sources (Lavrenko, 2000; Ogilvie, 2000; Lavrenko, 2002), and accommodating

generation of common words in a query (Zhai & Lafferty, 2001b). A number of studies have shown that smoothing has a very strong impact on the performance of LM-based retrieval systems (Ponte, 2001; Zhai & Lafferty, 2001a).

There are many smoothing techniques that have been proposed, mostly for speech recognition tasks (Chen & Goodman, 1998). These include: 1) correcting the ML estimates by pretending each n-gram occurs δ times more than it does with $0 < \delta \leq 1$ (Laplace, 1995; Lidstone, 1920; Johnson, 1932; Jeffreys, 1948); 2) discounting or scaling all non-zero ML estimates by a small constant amount and distributing the probability mass so gained uniformly over unseen events (Ney & Essen, 1993; Ney et al., 1994); 3) interpolating probability estimates from different models[5] (e.g. n-grams of different orders) provided that the contribution of each is weighted so that the result is another probability distribution (Jelinek & Mercer, 1980; Chen & Goodman, 1996); and 4) recursively backing off to lower order n-grams (Katz, 1987; Ney et al., 1994; Kneser & Ney, 1995). Note that back-off models of 4) can be thought of as a special case of the general linear interpolation model of 3) if the weights of the interpolated components are chosen so that their value is 0 except for the weight of the model that would have been chosen using the back-off method, which has the value 1.

As an alternative to ML, the Good-Turing estimate (Good, 1953) is often used (usually not by itself but in combination with other techniques) and has been central to many smoothing techniques. There are a number of other methods: the Witten-Bell smoothing (Bell et al., 1990; Witten & Bell, 1991) which can be viewed as an instance of interpolation-based methods, the Kneser-ney smoothing (Kneser & Ney, 1995) which is an extension of the discounting methods, and Bayesian smoothing with Dirichlet priors (MacKay & Peto, 1995) or beta prior (Nadas, 1984), to name a few. Detailed technical treatments of these methods can be found in (Jelinek,

[5] Another name for interpolation-based models is mixture models.

1997) and (Chen & Goodman, 1998). Among the smoothing strategies taken by IR researchers, many are borrowed from the speech community but there has also been considerable development where smoothing has been considered specifically for the IR domain. Different situations call for different strategies. We survey the strategies that have been used in LM for IR below.

Parameter smoothing. In LM approach to IR, one has to infer a language model $P(w|D)$ for each document D which is a probability distribution of all possible words w (e.g. in the vocabulary) in that document. Clearly, a document by itself is too small a sample for which to derive good ML estimates. To produce a reasonable estimated document language model, Ponte & Croft (1998) and Ponte (1998) employ a variety of smoothing techniques. For observed terms, their ML estimates are adjusted by a factor based on the average probability of a term in documents containing it and the associated risk (which is a geometric distribution). For unseen terms, the estimates simply back off to the collection probabilities. As an improvement to the basic method, a histogram estimator is used for low frequency terms.

The two-state HMM developed by Miller et al. (1999) can be viewed as a two-part mixture model. Although the motivations are different, the effect is to adjust the estimates of the query terms by using additional information from the collection.

Song & Croft (1999) use Good-Turing estimate in place of the ML estimate for the document model and then interpolate each document model with a background (or collection) model to produce the final document model (equation (4)). In addition, they consider term pairs and interpolate the unigram document model with a bigram model with the intuition that phrases would aid retrieval.

Differing from the above approaches in which the background model is estimated by using relative term frequencies in collection, Hiemstra (1998) use relative document frequencies of terms for smoothing. He argues that the background model so constructed is, in effect, a probabilistic interpretation of the traditional *idf* heuristics, and that the final smoothed document model achieves the effect of traditional *tf-idf* weighting with document length normalization.

In other work (Hiemstra, 2002), Hiemstra proposes to model the importance of the query terms explicitly. Previous approaches have treated each query term to be equally important, thus a single smoothing parameter λ is used for all query terms (see equations (3) and (4)). Hiemstra (2002) suggests using a different smoothing parameter λ_i for each query term q_i, thus the model becomes

$$P(Q \mid D) = \prod_{i=1}^{m} (\lambda_i P_{document}(q_i \mid D) + (1 - \lambda_i) P_{collection}(q_i))$$

Relevance feedback can be done with this model by boosting the λ values of the query terms that appear in the relevant documents while deemphasizing those terms that don't.

Jin et al. (2002) construct title language models based on the idea of statistical machine translation. While the use of translation models in their approach is theoretically motivated, it also plays a role of smoothing. Jin et al. (2002) first treat the titles as translations of documents and train a translation model based on the document-title pairs in the whole collection. Instead of being estimated only based on the observations of the document titles, the title language model of a document is now estimated by applying the learned translation model to the document. The translation model helps alleviate the data sparseness problem to some degree, but this is not enough. The training data is still sparse given the large number of parameters involved. To cope with this, Jin et al. (2002) extend the standard learning algorithms of the translation models by adding special parameters to model the self-translation probabilities of words.

Zhai & Lafferty (2001a) study a number of interpolation-based approaches to smoothing including Jelinek-Mercer smoothing, Bayesian smoothing with Dirichlet priors, and absolute discounting, as well as their back-off versions. Several large and small TREC collections are used for evaluating these methods. They find that different situations call for different approaches to smoothing, that retrieval performance is generally sensitive to smoothing parameters, and that the effect of smoothing is strongly dependent on the type of queries. They explain their empirical results by suggesting that smoothing plays two roles: one is to overcome the sparse data problem, and the other is to help generate common words in a query so as to cause query terms to be weighted in a similar fashion as the *idf* heuristics (Zhai & Lafferty, 2001b). Motivated by decoupling the two different roles of smoothing, Zhai & Lafferty (2002) develop a two-stage smoothing method. In the first stage, the document language model is smoothed with a Dirichlet prior with the collection model, and in the second stage, the smoothed document model is further interpolated with a query background model. An important advantage of this smoothing method is that it allows fully automatic estimation of the parameters. It is shown to be quite effective compared to a single-stage smoothing with an exhaustive parameter search on the test data.

Semantic smoothing. To move beyond parameter smoothing which plays a role similar to traditional term weighting, researchers have begun to look at semantic smoothing – another role that smoothing plays. As we pointed out earlier in this section, smoothing (more specifically linear interpolation or mixture models) can be used to combine knowledge sources. Most language models in IR can be generalized to the form of a mixture of a sparse topic model and a background model, and the smoothing strategies we have discussed so far approximate the topic model by counting words in a sample of text (e.g. document). What if we have more information

about the topic than just the document? The idea of semantic smoothing is to adjust the probability estimates of terms by exploiting context of words, for example, relevant documents, so that terms that are closely related to the original topic will get higher probabilities. The translation model proposed by Berger & Lafferty (1999) captures semantic relations between words based on term co-occurrences. However, the semantic relations between words are captured not by the language model but by the translation model, as they are learned from synthetic query/document pairs rather than the co-occurrences within the document collection. Hofmann (1999a, 1999b) discusses 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 (2001) propose a similar approach based on clustering.

Lavrenko (2000) hypothesizes that he could achieve semantic smoothing by using a *zone* of closely related text samples. He estimates a contextual model based on the texts in the *zone* and interpolates this contextual model with the document model and the background model. He chooses to use the Witten-Bell smoothing for estimating the weights of different models. The main problem with this approach is that retrieval performance is extremely sensitive to the number of text samples in the *zone*. A similar problem is encountered by Ogilvie (2000) when he smoothes document models with models of their nearest neighbors.

Lafferty & Zhai (2001a) and Lavrenko & Croft (2001) propose to use a weighted mixture of top-ranked documents from the query to approximate a topic model. Lafferty & Zhai (2001a) make use of a Markov chain method to assign weights to documents whereas Lavrenko & Croft (2001) employ an estimate of a joint probability of observing the query words together with any word in the vocabulary. Both methods have enabled large improvements in retrieval performance

over models that do not use semantic smoothing, but again they are very sensitive to the number of top-ranked documents that are used for probability estimation. Lavrenko (2002) explore the possibility of automatically finding the optimal subset of documents to construct an optimal mixture model. Two types of mixture models - set-based and weighted – are considered. He proves that it is not feasible to compute set-based optimal mixture models. For estimating weighted mixture models, a gradient descent procedure is proposed. Retrieval experiments indicate that the weighted mixture models are relatively insensitive to the number of top-ranked documents used in the estimation.

4. Comparisons with traditional probabilistic IR approaches

The language modeling approach has introduced a new family of probabilistic models to IR. Several researchers have attempted to relate this new approach to the traditional probabilistic IR approaches and compare their differences. The first workshop on language modeling and information retrieval held at Carnegie Mellon University from May 31 to June 1, 2001, has facilitated this discussion (Croft et al., 2001a).

Sparck Jones & Robertson (2001) examine the notion of relevance in the traditional probabilistic approach (PM) and the new language modeling approach (LM), and point out that two distinctions should be made between the two approaches. The first and what they call *surface* distinction is that while in both approaches, a good match on index keys between a document and a query implies relevance, relevance figures explicitly in PM but is never mentioned in LM. The second and what they find more important difference is that the underlying principle of LM is to identify *the* ideal document that generates the query rather than a list of relevant documents. Thus once this ideal document is recovered, retrieval stops. Because

of this, they argue that it is difficult to describe important processes such as relevance feedback in the existing LM approaches. Lafferty & Zhai (2001a, 2001b) and Lavrenko & Croft (2001) address these issues directly and suggest new forms of the LM approach to retrieval that are more closely related to the traditional probabilistic approach. Lafferty & Zhai (2001b) argue that in the traditional probabilistic approach proposed by Robertson & Sparck Jones (1976) a document could be thought of as generated from a query using a binary latent variable that indicates whether or not the document is relevant to the query. They show through mathematical derivations that, if a similar binary latent variable is introduced to LM, these two methods are on equal footing in terms of the relevance ranking principle and interpretation of the ranking process. However, this does not mean that PM and LM are just a reversion of each other. The differences go beyond a simple application of the Bayes' law. They point out that document length normalization is a critical issue in PM but it is not so in LM. Another difference is that in LM we have more data for estimating a statistical model than in PM which is the advantage of "turning the problem around". Both the risk-minimization framework suggested by Lafferty & Zhai (2001a, 2001b) and the relevance model suggested by Lavrenko & Croft (2001) move 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. Greiff (2001) suggests that the main contributions of LM to IR lie in the recognition of the important role of parameter estimation in modeling and the treatment of term frequency as the manifestation of an underlying probability distribution rather than as the probability of word occurrence itself. Zhai & Lafferty (2002) point out that traditional IR models rely heavily on *ad hoc* parameter tuning to achieve satisfactory

performance whereas in LM, statistical estimation methods can be used to set parameters automatically.

Hiemstra & de Vries (2000) relate LM to traditional approaches by comparing Hiemstra's model (1998) with the *tf-idf* term weighting and the combination with relevance weighting as done in the BM25 algorithm. They conclude that LM and PM have important similarities in that LM provides a probabilistic interpretation and justification of the *tf-idf* weighting and gives insight in why the combination of it with relevance weighting in BM25 is effective. Fuhr (2001) show how the LM approach can be related to other probabilistic retrieval models (Wong & Yao, 1995) in the framework of uncertain inference.

5. Applications

The language modeling approach was initially developed for *ad hoc* retrieval and found to be very effective. Soon after, successful applications of this approach to other retrieval tasks were reported, including relevance feedback, distributed IR, cross-lingual IR, quantification of query ambiguity, passage retrieval, and others.

Query ambiguity. Predicting query performance and thus dealing gracefully with ambiguous queries have long been an interest and challenge in information retrieval. An often-used example of ambiguous queries is "bank", where the context in which it appears could be financial institutions, river bank, or a flight maneuver, among others. Given such a query, IR systems need more user information than just the query words to resolve the ambiguity. One obvious solution is relevance feedback. That is, the retrieval system treats all queries (regardless of their degree of ambiguity) in the same way initially and then refines the document list based on user clarification on the initial list. However, relying on relevance feedback to solve the

problem may not always be realistic. A more effective alternative is to determine if each query is ambiguous and then ask the user specific questions about ambiguous queries. Vague queries will be handled differently than the clear ones from the beginning. Croft et al. (2001b), Cronen-Townsend & Croft (2002), and Cronen-Townsend et al. (2002) have developed a "clarity" measure within the language model framework to quantify query ambiguity with respect to a collection of documents without relevance information. In this approach, a query receives a non-negative "clarity" value based on how different its associated language model is from the language model of the collection, with a zero meaning that the query is maximally ambiguous whose associated language model is indistinguishable from the collection language model. Formally, the "clarity" measure is defined as the Kullback-Liebler (KL) divergence between the query distribution and the collection distribution (Croft et al., 2001b; Cronen-Townsend et al., 2002).

$$clarity \equiv KL(Q \,||\, Coll) = \sum_{w \in V} P(w \mid Q) \log \frac{P(w \mid Q)}{P_{coll}(w)}$$

where w is any term, Q is the query, $Coll$ is the collection, V is vocabulary of the collection, $P_{coll}(w)$ is the relative frequency of the term w in the collection, and $P(w|Q)$ is the query language model. Two types of language models, namely, the probability-weighted models and relevance models, have been used in their work for creating the query language model. In the probability-weighted approach, the query language model is taken to be

$$P(w \mid Q) \cong \sum_{D \in \mathrm{Re}t} P(w \mid D) P(D \mid Q)$$

where D is a document, Q is a query, and Ret is the set of retrieved documents. $P(D|Q)$ is the Bayesian inversion of $P(Q|D)$ with uniform document priors. Individual document models $P(w|D)$ are estimated using equation (8) and $P(Q|D)$ is estimated using equations (3) and (8). In

the relevance model approach, the query language model is given by $P(w \mid R)$ as defined in equations (10) and (11). Croft et al. (2001b) observe that the two types of query language models produce similar ranking when ordering queries based on their clarity scores. Cronen-Townsend & Croft (2002) and Cronen-Townsend et al. (2002) demonstrate that the clarity measure is highly correlated with the retrieval performance of the query and show how thresholds can be automatically determined to identify queries with poor language models.

Relevance feedback. The basic approach to relevance feedback has been to modify the query using words from top-ranked (for query expansion) or identified relevant documents. One way to use LM for this task is to build a language model for the top-ranked or relevant documents, and augment the query with words that have a relatively high log ratio of the probability of occurring in the model for relevant documents against the probability of occurring in the background (collection) model. This is the approach taken by Ponte (2000). In his work, the language model of the relevant documents is simply the sum of the individual document models. In the HMM system used by Miller et al. (1999), query expansion is achieved by adding to the initial query the words appearing in two or more of the top N retrieved documents and adjusting the HMM transition probabilities through training over queries. Croft et al. (2001b) view the query as a sample of text from a model of the information need. They hypothesize that users could also be represented as a mixture of topic language models generated from previous interactions and other sources. The task then boils down to estimating a language model associated with the query. Once we have the query language model, retrieval is straightforward – we can rank documents either according their likelihood of being generated by the query model (Lavrenko & Croft, 2001) or directly based on their similarity to the query model (Lafferty &

Zhai, 2001). Lavrenko & Croft (2001) develop relevance-based language models to approximate query models while Lafferty & Zhai (2001) make use of a Markov chain method.

Distributed IR. The major difference between distributed information retrieval and centralized retrieval is that distributed IR typically uses multiple collections at different sites so each site maintains its own index whereas centralized retrieval uses a centralized index for both indexing and retrieval. Therefore, one has to solve the problems of 1) resource selection, 2) *ad hoc* retrieval, and 3) results merging, when doing IR in a distributed environment. Language modeling can be used in any one of these steps or applied as a general integrated framework. Xu & Croft (1999) apply language modeling to resource selection. The basic idea is to group documents into clusters, each cluster representing a topic. A language model is built upon the word usage in a topic (cluster). To determine which topics are best for the query, the KL divergence is used to measure how well a topic model predicted the query. Based on this general idea, three different methods of organizing a distributed retrieval system are tested, all showing improvements over the traditional method of distributed retrieval with heterogeneous collections. Si et al. (2002) present an LM based framework that integrates all three sub-tasks of distributed IR. Query-based sampling is first applied to acquire language model descriptions of individual databases. In the resource selection step, databases are ranked according to the likelihood of a given query being generated from each of them. In the retrieval step, a language model based retrieval algorithm is used. In the result merging step, the document scores are recomputed so as to remove the possible bias caused by different collection statistics. The authors demonstrate through experiments that a simple language model can be used effectively in an intranet distributed IR environment.

Cross-lingual IR. The goal of cross-lingual IR is to find documents in one language for queries in another language. A straightforward adaptation of an LM approach to this task is to view the query in one language as generated from a document in another. Berger & Lafferty (1999) treat query generation as a translation process. Although their model has only been used for monolingual retrieval so far, it can easily accommodate a cross-lingual environment. Hiemstra & de Jong (1999) extend the model proposed in (Hiemstra, 1998; also given in equations (4)-(6)) to incorporating statistical translation for use in cross-lingual retrieval (in their work, Dutch queries on English document collections). By assuming the independence of the translation of a term and the document from which it is drawn, and also assuming each query term q_i has k_i possible English translations t_{ij} ($1 \leq j \leq k_i$), they end up with

$$P(Q \mid D) = P(q_1 \ldots q_m \mid D) = \prod_{i=1}^{m} \sum_{j=1}^{k_i} P(q_i \mid t_{ij}) P(t_{ij} \mid D)$$

where $P(t_{ij}|D)$ is computed using equations (4), (5), and (6). Note that the probability estimates of query terms in the source language are, in effect, smoothed by using the background model built on a document collection in the target language. Xu & Weischedel (2000) and Xu et al. (2001) suggest that if the background model can be built directly using a document collection in the source language then the noise introduced by translation can be avoided. Motivated by this, they present an extension of the monolingual HMM model proposed by Miller et al. (1999) to use for querying Chinese (or Spanish) document collections with English queries. The computation involves the following expression

$$P(Q_e \mid D_c) = \prod_{i=1}^{m} (\lambda P(q_i \mid GE) + (1-\lambda) \sum_{c \in Chinese} P(c \mid D) P(q_i \mid c))$$

where Q_e stands for an English query, D_c stands for a Chinese document, *GE* stands for general English, and c is any word in Chinese. Their cross-lingual system achieves roughly 90% of

monolingual performance in retrieving Chinese documents and 85% in retrieving Spanish documents. Lavrenko et al. (2002) apply the relevance model to a similar task with new estimation methods that are adapted to cross-lingual retrieval. The model starts with a query in the source language and directly estimates the model of relevant documents in the target language, which is different from other models in that it does not attempt to translate either documents or the query. Comparable retrieval performance to other models is observed.

Passage retrieval. Given that early work on language modeling for IR has been entirely document-based, Liu & Croft (2002) address the question whether language models would be feasible to use for passage retrieval. In their work, a probability-weighted query likelihood model and a relevance model are used. After experimenting with different types of passages (e.g. half-overlapped windows and arbitrary passages at different passage sizes) and various ways of constructing the language models for doing passage retrieval, they find that all passage retrieval runs produce comparable results with and sometimes significant improvements over full-length document retrieval when using language models. A second finding is that passage retrieval can provide more reliable performance than full-length document retrieval in the language model framework, especially when using relevance models.

Incorporating prior knowledge. It is common in IR that documents are treated equally likely to be relevant, that is, the prior probability *P(D)* is assumed uniform (see also equation (1)). However, in practice, some documents that have certain properties may be more likely to be relevant given the task at hand. An example of such properties is document length. If these properties are known and can be exploited by the retrieval system, they should help improve retrieval performance. Traditional IR approaches have attempted to incorporate prior knowledge, but the incorporation was rather heuristic and not based on formal models. The LM approach to

IR, in contrast, has provided a natural and principled way of incorporating knowledge. Several papers discuss their efforts in using prior knowledge with the LM framework. Hiemstra & Kraaij (1999) show that the document length is helpful for *ad hoc* retrieval, especially for short queries. Miller et al. (1999) combine several features in their document priors, including document source, length, and average word-length. Small improvement over uniform priors is observed. Zhu (2001) discuss how to derive features beyond bag of words in language models via discriminative techniques. Kraaij et al. (2002) use several web page features such as the number of inlinks and the form of the URL as prior knowledge for an LM-based entry page retrieval system. They show that language models can accommodate such knowledge in an almost trivial manner, by estimating the posterior probability of a page given its features and using this posterior probability as the document prior in the language model. Evaluations demonstrate that significant performance improvements over the baseline LM system can be obtained through proper use of prior probabilities. Li & Croft (2003) show that time serves as a useful prior for queries that favor very recent documents and those that have more relevant documents within a specific period in the past. Depending on the type of query, either an exponential distribution or a normal distribution is used to replace the uniform prior probability in a query likelihood model and a relevance model.

6. Research directions

Statistical language modeling has brought many opportunities to IR. First of all, it provides us with a formal method of reasoning about the design and empirical performance of an IR system. Many heuristic techniques introduced in the past can now be explained and accommodated by this new framework. A case in point is the *tf-idf* weighting used in traditional

IR systems. Second, since language models have been in use for nearly thirty years in other language processing tasks such as automatic speech recognition, lots of experience has been gained and lessons learned that could be leveraged by the IR community. For example, smoothing has been studied extensively in the speech community and many smoothing methods have been developed. Researchers in the IR community have started looking at what is available there and either adapting existing ones to retrieval tasks or developing new techniques based on them. Third, the statistical language modeling approach applies naturally to a variety of information system technologies, such as *ad hoc* and distributed retrieval, cross-language IR, summarization and filtering, topic tracking and detection, and, possibly, question answering. Preliminary successes have been reported in many of these areas. Because of the above advantages, LM is considered the most promising framework for advancing information retrieval to meet future challenges (Allan et al., 2003).

One long-term challenge of IR is to provide global information access to users such that information needs expressed in any language can be satisfied by information from many sources, encoded in any language. This calls for developments of massively distributed, cross-lingual retrieval systems. While existing techniques in distributed IR, cross-lingual IR, and possibly data fusion can easily lend themselves to the design of such a system, simply combining these techniques would not result in a satisfactory solution. Therefore, to cope with this challenge, current LM methods must be extended to provide a unified retrieval framework that leverages techniques from multiple areas/fields.

Another long-term challenge of IR is to explicitly capture information about the user and context of the information retrieval process, and integrate models of users into the retrieval models. In current IR systems, little has been done to achieve this and models of the user are

weak if not missing from retrieval models. As a result, current IR systems resort to a user-generic approach in which different users with the same query will be provided with the same results. While this approach has proven to be good enough for an average user, it leaves much room for improvements in retrieval effectiveness for individual users. One hope for LM to help IR move beyond this is to represent the user by a probability distribution of interests (e.g. words, phrases, or topics), actions (e.g. browsing behavior), and annotations/judgments (Allan et al., 2003). Knowledge about the user-related context (e.g. user type, background, and personalization) and task-related context (e.g. genre, level, authority, and subject domain) can be encoded in the priors of the language models.

As stepping-stones towards achieving the long-term goals, there are some shorter-term challenges that LM must cope with. Current LM techniques must be extended to incorporate diverse data sources and multiple forms of evidence. In web search, for instance, it is a challenge to exploit all sorts of evidence, including the web structure, meta-data, user context information, to find high quality information for users. This calls for LM to allow more precise representation of information needs, integration of structural evidence, and incorporation of linguistic/semantic knowledge. For question answering (QA), the standard approach has been locating documents or passages with candidate answers, and then integrating multiple passages and multiple data sources, including structured and unstructured text, to provide the final answer. Current statistical language models have proven useful for the document/passage retrieval part, but are not adequate for finding the exact answers. A promising direction seems to be to extend language models to include structured patterns which are used rather heuristically in existing systems.

New LM techniques must also be developed to support more advanced retrieval tasks as well as provide more integrated models for current tasks. For example, retrieval would benefit

from adding structure (e.g. proximity operators) to the query. The challenge for LM is to study how the structure can be represented in probabilistic terms. In cross-lingual IR, most current LM approaches make use of a probabilistic model for translation and a language model for retrieval. However the two components are only loosely coupled and independence assumptions are made for both of them. One possible way to improve this is to explore models with loosened independence assumptions, thus enabling the direct use of contextual information of words and phrases for translation. LM techniques that allow for a tighter integration of the translation model and the retrieval model are likely to be explored. In addition, better mechanisms for relevance feedback may be necessary. For distributed IR, while techniques developed for *ad hoc* retrieval have been successfully employed, it would be more desirable to develop a theoretically grounded model of its own, in hope of providing a unified model for the metasearch problem (data fusion and collection fusion).

As LM for IR grows, existing models will continue to be refined, parameter estimation procedures will be improved, and the relationships between the various modeling approaches will be examined more carefully. Interest in applying more smoothing techniques and combining language models for multiple topics and collections is likely to continue. For example, the probabilistic latent semantic indexing technique (Hofmann, 1999) seems to be a promising technique to be used for topic models. Models based on clustering techniques may also be explored. We can expect more sophisticated language modeling techniques to be developed that allow for increasingly integrated representations across documents, collections, languages, topics, queries, and users.

Acknowledgments

This review was supported in part by the Center for Intelligent Information Retrieval, in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903, and in part by Advanced Research and Development Activity under contract number MDA904-01-C-0984. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

References

Allan, J. (editor), Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, W.B. (editor), Dumais, S., Fuhr, N., Harman, D., Harper, D., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J., & Zhai, C. (2003). Challenges in information retrieval and language modeling. *SIGIR Forum*, vol. 37, no. 1.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press Series/Addison Wesley, New York,

Bell, T. C., Cleary, J. G., & Witten, I. H. (1990). *Text Compression*. Englewood Cliffs, NJ: Prentice Hall.

Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference*, pp. 222-229.

Chen, S. F. & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 310-318.

Chen, S. F., & Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Center for Research in Computing Technology, Harvard University, 1998.

Cooper, W. S. (1995). Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval, *ACM Transactions on Information Systems (TOIS)*, vol.13 no.1, pp.100-111.

Cooper, W. S., & Maron, M. E. (1978). Foundations of probabilistic and utility-theoretic indexing, *Journal of the ACM (JACM),* vol. 25, no.1, pp.67-80.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

Croft, W. B., Callan, J., & Lafferty, J. (2001a). Workshop on language modeling and information retrieval. *SIGIR Forum*, vol. 35, no. 1.

Croft, W. B., Cronen-Townsend, S., & Lavrenko, V. (2001b). Relevance feedback and personalization: A language modeling perspective. In *Proceedings of the DELOS-NSF Worshop on Personalization and Recommender Systems in Digital Libraries*, pp. 49-54.

Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of documentation*, 35, pp. 285-295.

Croft W. B., & Lafferty, J (eds.) (2003). *Language Modeling for Information Retrieval*. Kluwer, 2003.

Cronen-Townsend, S., & Croft, W. B. (2002). Quantifying query ambiguity. In *Proceedings of the Human Language Technology 2002 Conference*, pp. 94-98.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, pp. 299-306.

Fuhr, N. (1989). Models for retrieval with probabilistic indexing. *Information Processing and Management*, vol. 25, no. 1.

Fuhr, N. (2001). Language models and uncertain inference in information retrieval. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, 2001.

Good, I. J. (1950). *Probability and the Weighting of Evidence*. Charles Griffin and Co. Ltd., London.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40, pp. 237-264.

Greiff, W. (2001). Is it the language model in language modeling? In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, Pittsburgh.

Hiemstra, D. (1998). A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the Second European Conference on Research and Advance Technology for Digital Libraries (ECDL)*, pp. 569-584.

Hiemstra, D. (2002). Term-Specific Smoothing for the Language Modeling Approach to Information Retrieval:The Importance of a Query Term. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, pp 35-41.

Hiemstra, D., & de Jong, F. (1999). Disambiguation strategies for cross-language information retrieval. In *Lecture Notes in Computer Science: Research and Advanced Technology for Digital Libraries* (vol. 1696), Springer-Verlag, pp. 274-293.

Hiemstra, D., & de Vries, A. (2000). Relating the new language models of information retrieval to the traditional retrieval models. In *CTIT Technical Report* TR-CTIT-00-09.

Hiemstra, D., & Kraaij, W. (1999). Twenty-One at TREC-7: Ad-hoc and cross-language track. In E.M. Voohees, & D.K. Harman (Eds.) *The Seventh Text Retrieval Conference (TREC7)*, NIST.

Hofmann, T. (1999a). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference*.

Hofmann, T. (1999b). From latent semantic indexing to language models and back. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, Pittsburgh.

Jeffreys, H. (1948). *Theory of Probability*. Oxford: Clarendon Press.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, Massachusetts.

Jelinek, F., & Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice,* Amsterdam, The Netherlands: North-Holland, May.

Jin, R., Hauptmann, A. G., & Zhai, C. (2002). Title Language Model for Information Retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, pp. 42-47.

Johnson, W. E. (1932). Probability: deductive and inductive problems. *Mind,* 41, pp. 421-423.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing,* ASSP-35(3), pp. 400-401, March.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Volume I, pp. 181-184, Detroit, Michigan.

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In M. Beaulieu, R. Baeza-Yates, S.H. Myaeng, & K. Järvelin (Eds.), *Proceedings of the 25th annual international ACM-SIGIR conference on research and development in information retrieval*, Tampere, Finland, pp.27-34, New York: ACM.

Lafferty, J. & Zhai, C. (2001a). Document language models, query models, and risk minimization for information retrieval. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), *Proceedings of the 24th annual international ACM-SIGIR conference on research and development in information retrieval*, New Orleans, Louisiana (pp.111-119), New York: ACM.

Lafferty, J., & Zhai, C. (2001b). Probabilistic IR models based on document and query generation. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, Pittsburgh.

Laplace, Pierre Simon marquis de. (1995). *Philosophical Essay on Probabilities*. New York: Springer-Verlag.

Lavrenko, V. (2000). Localized smoothing of multinomial language models. *CIIR Technical Report IR-222*, University of Massachusetts, Amherst.

Lavrenko, V. (2002). Optimal Mixture Models in IR. In *Proceedings of the 24th European Colloquium on IR Research (ECIR'02)*, Glasgow, Scotland, March 25-27, 2002.

Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana (pp.120-127), New York: ACM.

Lavrenko, V., Choquette, M., & Croft, W. B. (2002). Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*.

Lidstone, G. J. (1920). Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8, pp. 182-192.

Li, X., & Croft, W. B. (2003). Time-based language models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM'03)*.

Liu, X., & Croft, W. B. (2002). Passage Retrieval Based On Language Models. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02)*, pp. 375-382.

MacKay, D. J. C., & Peto, L. C. B. (1995). A hierarchical Dirichlet language model. *Natural Language Engineering*, vol. 1, no. 3, pp. 1-19.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Language Processing*. The MIT Press, Cambridge, Massachusetts.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7, pp. 216-244.

Miller, D., Leek, T., & Schwartz, R. (1999). A hidden Markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference*, pp. 214-221.

Nadas, A. (1984). Estimation of probabilities in the language model of the IBM speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing,* ASSP-32(4), pp. 859-861, August.

Ney, H., & Essen, U. (1993). Estimating 'small' probabilities by leaving-one-out. In *Eurospeech'93*, vol. 3, pp. 2239-2242, ESCA.

Ney, H., Essen, U., & Kneser, R. (1994). On structuring probabilistic dependences in stochastic language modeling. *Computer Speech and Language*, 8:1-38.

Ogilvie, P. (2000). Nearest neighbor smoothing of language models in IR. *Unpublished work.*

Peters, J. (2001). Semantic text clusters and word classes – the dualism of mutual information and maximum likelihood. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, Pittsburgh.

Ponte, J. (1998). *A language modeling approach to information retrieval.* Ph.D. dissertation, Dept. of Computer Science, University of Massachusetts, Amherst.

Ponte, J. (2000). Language Models for Relevance Feedback. In W. B. Croft (Ed.), *Advances in Information Retrieval: Recent Research from the CIIR.* Kluwer Academic Publishers, chapter 3, pp. 73-95.

Ponte, J. (2001). Is information retrieval anything more than smoothing? In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, Pittsburgh.

Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM-SIGIR conference on research and development in information retrieval*, pp.275-281, New York: ACM.

Robertson, S. E. (1977). The probability ranking principle in IR, *Journal of Documentation*, 33, pp. 294-304.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms, *Journal of the American Society for Information Science*, vol. 27, no. 3, May-June, pp. 129-146.

Robertson, S. E., Walker, S., Sparck Jones, K., Hancock-Beaulieu, M. M., & Gatford, M, (1995). Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)* edited by D.K. Harman. NIST Special Publication 500-225.

Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? In *Proceedings of the IEEE*, 88(8), 2000.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, vol. 30, pp. 50-64.

Si, L., Jin, R., Callan, J., & Ogilvie, P. (2002). Language modeling framework for resource selection and results merging. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02)*.

Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.279-280, New York: ACM.

Sparck Jones, K. & Robertson, S. (2001). LM vs. PM: where is the relevance? In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, Pittsburgh.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000a). A probabilistic model of information retrieval: development and comparative experiments - Part 1. *Information Processing and Management*, vol. 36, no. 6, pp. 779-808.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000b): A probabilistic model of information retrieval: development and comparative experiments - Part 2. *Information Processing and Management*, vol. 36, no. 6, pp. 809-840.

Turtle, H. R. (1990). *Inference networks for document retrieval*. Ph.D. dissertation, University of Massachusetts, Amherst.

Turtle, H., & Croft, W. B. (1991). Efficient probabilistic inference for text retrieval. In *Proceedings of RIAO 3*.

van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, vol. 33, no. 2, pp.106-119.

van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths.
http://www.dcs.gla.ac.uk/Keith/Preface.html

van Rijsbergen, C. J. (1992). Probabilistic retrieval revisited. *The Computer Journal*, vol. 35, no. 3, pp. 291-298.

Witten, I. H., & Bell, T. C. (1991). The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1085-1094, July.

Wong, S., & Yao, Y. (1995). On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, vol. 13, no. 1, pp. 38-68.

Xu, J., & Croft, W. B. (1999). Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.254-261, New York: ACM.

Xu, J., & Weischedel, R. (2000). TREC-9 cross-lingual retrieval at BBN. In E.M. Voorhees, & D.K. Harman (Eds.), *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, Gaithersburg, MD.

Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual retrieval. In *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 105-110.

Zhai, C. (2002). *Risk Minimization and Language Modeling in Text Retrieval*. Ph.D. dissertation, Carnegie Mello University.

Zhai, C., & Lafferty, J. (2001a). A study of smoothing methods for language models applied to ad hoc information retrieval. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana (pp. 334-342), New York: ACM.

Zhai, C., & Lafferty, J. (2001b). Dual role of smoothing in the language modeling approach. In *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR) 2001*, pp. 31-36, Pittsburgh.

Zhai, C., & Lafferty, J. (2002). Two-stage language models for information retrieval. In *Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 49-56.

Zhu, J. X. (2001). Language model feature induction via discriminative techniques. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, Pittsburgh.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, vol. 40.

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison Wesley Press.

Zipf, G. K. (1965). *The Psycho-Biology of Language: An introduction to Dynamic Philology*. MIT Press, Cambridge.