Language Modeling Approaches to Information Retrieval

Protima Banerjee
College of Information Science & Technology, Drexel University,
3141 Chestnut Street, Philadelphia PA 19104
protima.banerjee@drexel.edu

Hyoil Han
College of Information Science & Technology, Drexel University,
3141 Chestnut Street, Philadelphia PA 19104
hyoil.han@acm.org

Received 3 March 2009; Accepted 8 April 2009

This article surveys recent research in the area of language modeling (sometimes called statistical language modeling) approaches to information retrieval. Language modeling is a formal probabilistic retrieval framework with roots in speech recognition and natural language processing. The underlying assumption of language modeling is that human language generation is a random process; the goal is to model that process via a generative statistical model. In this article, we discuss current research in the application of language modeling to information retrieval, the role of semantics in the language modeling framework, cluster-based language models, use of language modeling for XML retrieval and future trends.

Categories and Subject Descriptors: Information Retrieval and Visualization, Natural Language Processing

General Terms: Language Modeling, Statistical Language Modeling, Information Retrieval

Additional Keywords and Phrases: Semantic Smoothing, Cluster-based Language Models, XML Retrieval

1. INTRODUCTION

Language modeling is a formal probabilistic retrieval framework with roots in speech recognition and natural language processing [Jurafsky and Martin 2000]. The underlying assumption of language modeling is that human language generation is a random

process; the goal is to represent that process via a statistical model. Using a language model, we can calculate the likelihood of a language sequence, such as a sentence, being generated.

Language models were first successfully applied to information retrieval by [Ponte and Croft 1998]. In that work, the authors proposed a query-likelihood model [Liu and Croft 2005] in which a query is considered to be generated from an "ideal" document that satisfies an information need. The retrieval engine estimates the likelihood that each document in the corpus is the ideal document, and then ranks the documents accordingly. The underlying premise to this approach is that each document in the corpus has a different language model [Manning et al. 2007]. This allows the use of statistical techniques to both estimate document models and to score documents against a particular query.

Later works [Song and Croft 1999; Lavrenko and Croft 2001; Zhai and Lafferty 2004] expand on this early research using more sophisticated models that include topics, phrases and relevance. All confirm that language modeling techniques are preferred over *tf-idf* (term frequency-inverse document frequency) weights [Robertson and Jones 1997], because of empirical performance as well as the probabilistic meaning that can be formally derived from a language modeling framework. In contrast to the classic vector space model [Salton and McGill 1986] which produces a geometric document score, a language model produces a likelihood estimate which is intuitively easier to understand. The majority of language modeling approaches to information retrieval can be categorized into one of four groups: (1) the generative query likelihood approach, which ranks based on the likelihood of a document language model generating the query, (2) the generative document likelihood approach, which ranks based on the likelihood of a query language model generating a document, (3) the comparative approach, which ranks based on the similarity between the query language model and document language model, (4) translation models, which rank based on the likelihood of the query being viewed as a translation of a document, and (5) cluster-based language models. The remainder of this paper will describe the foundations of statistical language modeling and trace the research that has been done in each of these categories. In addition, we will also describe extensions to language modeling that have been developed for XML retrieval and how semantics can be incorporated into the language modeling framework. This paper will conclude with a discussion on why language modeling approaches continue to be an active area in information retrieval research, and future trends for that body of work.

2. STATISTICAL LANGUAGE MODELING

The roots of statistical modeling date back to the middle of the twentieth century with the work of Shannon [1951]. Shannon proposed a conceptual framework in which a subject was asked to guess the next letter in a sequence from a stream of printed English, given the preceding letters and words in the passage. The subject is informed if he guessed correctly; if he guessed incorrectly, he is told the correct letter in the sequence and can then proceed to the next letter. Only the letters that are incorrectly guessed, that is, the letters where the system must provide the correct letter to the subject, are written down. The next subject in the experiment would then be asked

to re-construct the original sequence from the stream of incorrectly guessed letters, which Shannon calls the "reduced text." The underlying premise of this experiment, and the concept which carries through to statistical language modeling, is that a stripped down or lossy version of a full text can be used to reconstruct the complete document. This model is very applicable to fields such as speech recognition, where there is typically some loss in fidelity between a steam of spoken words and its computer transcription to text.

The work of [Zipf 1949] is also fundamental to statistical language modeling. While studying statistical word occurrences in natural language texts, Zipf postulated that the frequency of word occurrences in a document is roughly inversely proportional to a word's rank within the document, if the words are sorted by occurrence frequency. Conceptually, Zipf's Law says that "while only a few words are used very often, many or most are used rarely [Zipf 1949]." Zipf's work evidences the applicability of statistical models to natural language; it should be noted that while his most prominent work dealt with English language texts, he also successfully applied the same model to languages such as Chinese, which have a widely divergent grammar from English.

Fundamentally, a statistical language model is a generative pattern of language; that is, it seeks to estimate the probability of occurrence of a sequence of words. Language modeling is predictive in nature; the goal of the model is to estimate the probability of future words, given the pattern of words that we already know [Manning and Schütze 1999]. Language modeling has been applied to a variety of fields including natural language processing, speech recognition, machine translation, and information retrieval. A full survey of statistical language modeling techniques across these divergent applications is beyond the scope of the current paper, and the reader is encouraged to review Rosenfeld's survey on the topic [Rosenfeld 2000] for an overview. Seminal papers in the field have included the work of [Bahl et al. 1989], which discusses statistical language modeling in the context of speech recognition, [Jelinek and Mercer 1980], which presents a method for estimating language model parameters in a sparse data environment and Katz's method [Katz 1987] for CPU and storage efficient language model computation.

The perplexity metric is commonly used to assess the effectiveness of a language model [Jelinek et al. 1977]. Perplexity is a type of information entropy measure; mathematically, perplexity is defined by the formula below, where H is the entropy measure for a particular random variable X having a probability distribution $p(x)$.

$$perplexity = 2^{H(p)} = 2^{-\Sigma p(x)\log 2 p(x)} \tag{1}$$

Conceptually, perplexity can be thought of as a confidence measure of the predictive properties of the model. At a word or passage level, the perplexity measure defines how confidently we can define the next word in a sequence given the context of the previous words. Language models can be evaluated against one another by comparing their perplexity measures. It should be noted that perplexity is domain dependent; a given language model will typically have much lower perplexity (better performance) in a highly specialized domain than in general English.

The simplest form of language model is the "bag of words" or unigram model, which examines each word independently of its context [Manning et al. 2007]. There are

many other more complex types of language models such as bi-gram models or tri-gram models which condition the existence of the next word based on the previous word or two words; words are considered as sets of two or three, respectively in these approaches [Jurafsky and Martin 2000]. Many modern approaches [Rosenfeld 2000] have used large corpora to train language models; these approaches have typically used simple models such as unigram or bi-gram approaches to language models, relying on the large quantities of training texts of various types to improve performance.

The fundamental problem of language modeling is that we never have a clear confirmation of the specific model that we are assigning to any given document [Manning et al. 2007]. This is because the language model is an estimate of word occurrence probabilities based on the text of a document; we treat each document as a piece of representative text from its underlying language model. Thus, the creation of a language model for a given document is essentially the problem of deriving a complete word occurrence probability model from an incomplete or lossy sample (the document text). In this situation, there are often cases where words which ought to be included as a part of the language model for a document are, in fact, not present in the actual text, even if the size of the text is huge. For example, a document about healthful eating ought to have a high probability of occurrence for the words "diet" and "nutrition" as a part of its language model even if those words are not in the text of the document. However, a document about healthful eating is less likely to contain the word "alligator" since it is less likely that a document about healthful eating will be related to alligators. From these examples one can see that it is problematic to use estimates such as the Maximum Likelihood Estimate which are based strictly on term counts for language model construction. This is often referred to as the "zero frequency problem" or the "sparse data problem [Witten and Bell 1991]."

Smoothing is a set of techniques used to address the "sparse data" issue. Smoothing is so called because it attempts to raise low or zero word-occurrence probabilities and lower high word-occurrence probabilities, based on prior knowledge about a passage, document or corpus [Zhai and Lafferty 2004]. Smoothing is a fundamental part of the language modeling paradigm; "in the language modeling approach, the accuracy of smoothing is directly related to performance [Zhai and Lafferty 2004]." In the sections below the smoothing techniques which are applied by each of the language modeling approaches will be explicitly described.

3. LANGUAGE MODELING APPROACHES TO INFORMATION RETRIEVAL

Since the introduction of language modeling to information retrieval (IR) by Ponte and Croft [Ponte and Croft 1998], there have been a number of IR approaches based on language modeling techniques As stated earlier, this paper categories IR research using language modeling techniques into four broad categories: (1) generative query-likelihood models, (2) generative document-likelihood models, (3) model comparison approaches, (4) statistical translation methods and (5) cluster-based language models. The remainder of this section explores the research that has been performed in each of these categories in detail.

3.1 Query-Likelihood Models

The premise of the query-likelihood approach is that each document in a collection can be thought of as having an individual document language model, and document language models within the collection can be ranked by their probability of being able to generate a particular query. The earliest work in the query-likelihood family of approaches can be considered to be that of Kalt [Kalt 1996]. Kalt considered that term probabilities for documents related to a single topic can be modeled by a single stochastic process; documents related to different topics would be generated by different stochastic processes. Kalt's model treats each document as a sample from a topic language model. Since the problem he considered was text classification, queries were derived from a training set rather than traditional query strings. Kalt's approach was based on the Maximum Likelihood Estimate (MLE) [Manning and Schütze 1999], and incorporated collection statistics, term frequency and document length as integral parts of the model. Although later query-likelihood approaches are more robust in that they consider that each document (vs. a group of documents) is described by an underlying language model, Kalt's early work is clearly a pre-cursor to language modeling in information retrieval.

The Ponte and Croft query-likelihood model [Ponte and Croft 1998] assumes a unigram language model and, like Kalt's model, starts from the basis of the MLE. Mathematically, MLE is defined below where $p(t \mid d)$ is the probability of a term given a document d, tf is the term frequency of the given term t, and dl is the total number of terms within the document, which can also be thought of as the document length. Ponte and Croft equate MLE to the probability of the term given the document's language model, or $p(t \mid M_d)$.

$$p(t \mid M_d) = \frac{tf}{dl} \tag{2}$$

To smooth the zero probability terms in the MLE, the authors take the background probability of all terms in the collection into account; that is, they augment the probability of the term appearing in any specific document with the probability of the term across the entire document collection. However, as not all the documents in the collection come from the same language model, a geometric risk function is incorporated into the smoothing methodology. This risk function essentially minimizes the impact of documents which might be "outliers" for a given term; in other words, it reduces the impact of those documents which have term frequencies which diverge from the normalized mean of the collection by reducing their contribution to the language model. Ponte and Croft incorporate the risk function into their approach by using it to modulate the influence of the term frequency within the specific document as well as the average occurrence of the term within the corpus.

In a later work, [Song and Croft 1999] apply a more generic view to the language modeling problem based on the same query-likelihood viewpoint; they evaluate a number of approaches to language modeling and propose a set of improvements that can either be used independently or in conjunction with one another. Specifically, they focus on smoothing document language models with the Good-Turing estimate [Manning and Schütze 1999] and curve fitting, expanding a document model with the

corpus document model, modeling a query as a sequence (rather than a set) of terms, and finally combining the unigram language model with a bi-gram language model.

The Good-Turing estimate adjusts the raw term frequency scores *tf* in the following manner:

$$tf = (tf+1)\frac{E(N_{tf+1})}{E(N_{tf})} \tag{3}$$

Here, N_{tf} represents the number of terms that have term frequency *tf*, and $E(N_{tf})$ is the expected value of the number of terms that have a term frequency of *tf*. Intuitively, the Good-Turing estimate states that the ratio of two adjacent term frequencies will be equivalent to the ratio of the expected values of the number of terms that have those frequencies. Practically, however, it is difficult to determine the number of terms that have specified frequencies due to the limited data available within a document; too many terms may have frequencies close to zero to make the Good-Turing estimate useful. Song and Croft adopt a curve fitting approach that uses a geometric distribution with a nested logarithmic exponent to approximate $E(N_{tf})$.

Song and Croft use the Good-Turning estimate to create a smoothed language model for both the individual document as well as the corpus. Then, they apply a weighted sum to combine the document language model with the corpus language model. The weighted sum approach is represented below:

$$p(t \mid d) = wP_{document}(t \mid d) + (1-w)P_{corpus}(t) \tag{4}$$

The weighting parameter w is a number between 0 and 1, and the weighted sum approach of combining probabilities has the advantage of always producing a normalized result; in other words, the weighted sum approach will always produce a number between 0 and 1 for $p(t \mid d)$.

Interpolation is also used to combine the unigram document model with a bi-gram document model. This is represented below:

$$p(t_i,t_{i-1} \mid d) = \lambda_1 p(t_i \mid d) + \lambda_2 p(t_i,t_{i-1} \mid d) \tag{5}$$

Here, the weighting parameters λ_1 and λ_2 should be set so that $\lambda_1+\lambda_2$ is equal to 1 for every term t. This is done empirically by Song and Croft; however, the Expectation Maximization algorithm [Dempster et al. 1977] can be used to set these parameters using a training corpus.

In terms of query processing, Song and Croft [1999] treat the query as a sequence of terms, as opposed to the set of terms approach adopted by Ponte and Croft [1998]. This can be represented as follows:

$$P_{sequence}(Q \mid d) = \prod_i p(t_i \mid d) \tag{6}$$

[Hiemstra 1999] presents a related approach which applies statistical language modeling to information retrieval. Hiemstra's approach emphasizes the importance of the ordering of the terms in the document. That is, "the most important modeling assumption we make is that a document and a query are defined by an ordered sequence of words and terms [Hiemstra 1999]." In this framework, text is modeled as

an ordered sequence of n random variables, one for each unique term which appears in the document. This is represented by the following models for documents and queries:

$$P(T_1,T_2,..., T_N \mid D) = \prod_i P(T_i \mid D) \text{ and } P(T_1,T_2,..., T_N \mid Q) = \prod_i P(T_i \mid Q) \quad (7)$$

where D and Q are the documents and queries, respectively, and T_i is the event that term i occurred in document D or query Q. The matching process between a query and a document are represented by the following formula:

$$P(D \mid Q) = \sum_\tau P(\tau \mid Q)P(D \mid \tau),\ \tau = T_1,T_2,..., T_N \quad (8)$$

Here τ represents the set of all possible term sequences T_1, T_2,..., T_N.

Hiemstra also addresses the sparse data problem. "We believe," he states, "that the sparse data problem is exactly the reason that it is hard for information retrieval systems to obtain high recall values without degrading values for precision." Hiemstra smoothes the probability distribution for each term using a linear interpolation of term frequency and document frequency. This is represented by the equation below:

$$P(T_i = t_i \mid d) = \alpha_1 \frac{df(t_i)}{\sum_t df(t)} + \alpha_2 \frac{tf(t_i,d)}{\sum_t tf(t,d)} \quad (9)$$

Here, $df(t_i)$ is the number of documents in which term t_i appears, which is also known as the document frequency and $tf(t_i,d)$ is the number of times term t_i that appears in document d, which is also known as the term frequency. The document frequency component (i.e., the first component) in the equation above can be thought of as the smoothing contribution coming from the document corpus, while the second component can be thought of as the contribution coming from an individual document. α_1 and α_2, are the weighting coefficients for the document frequency and term frequency contribution to the smoothing model, respectively. Hiemstra does not specifically proscribe a method for setting α_1 and α_2. "In general," he says, "one wants to find the combination of weights that works best, for example, by optimizing them on a test collection consisting of documents, queries, and corresponding relevance judgments [Hiemstra 1999]." It should be noted that this smoothing approach can be directly related to the well-known *tf-idf* approach.

Hiemstra's smoothing model is similar to Song and Croft's smoothing model in that both approaches are fundamentally a linear combination of a corpus document model (Hiemstra's document frequency component) and an individual document model (Hiemstra's term frequency component). However, there is one fundamental differences between the two approaches: Song and Croft smooth the document and corpus language model (via the Good-Turing estimate) prior to the linear interpolation; Hiemstra combines the two models without smoothing [Song and Croft 1999].

A two-stage Hidden Markov Model forms the basis for the query-likelihood language model presented by [Miller et al. 1999]. A discrete Hidden Markov Model is defined by "a set of output symbols, a set of states, a set of probabilities for transformations between the states, and a probability distribution on output symbols for each state [Miller et al. 1999]." The Hidden Markov Model is referred to as "hidden" because the

state transition process itself is never directly observed; it can only be inferred based on observed events. Hidden Markov Models are widely used in natural language processing and speech recognition [Jurafsky and Martin 2000].

The first stage in the proposed two-stage Hidden Markov Model represents the probability that a given query term will be found within the document; the second stage in the proposed Hidden Markov Model represents the probability that a given query term will be found within General English but is unrelated to the document The proposed model makes the simplifying assumption that a given query term will move to either the first or second stage; thus only two stage changes are considered to be part of the model.

[Miller et al. 1999] make simplifying assumptions similar to [Ponte and Croft 1998] and [Hiemstra 1999]. Rather than using the Expectation Maximization (EM) algorithm [Dempster et al. 1977] to compute transition probabilities and output distributions, the simpler MLE (Maximum Likelihood Estimate) method is used. In this model, the "General English" stage of the Hidden Markov Model provides a smoothing function. This stage is approximated by using the entire document corpus as an approximation for the full English language.

Mathematically, the two-stage Hidden Markov Model is represented by the following equation:

$$P(Q \mid D) \prod_{q \in Q} a_0 P(q \mid GE) + a_1 P(q \mid D) \tag{10}$$

where $P(Q \mid D)$ is the likelihood of the query being a representation of the document, $P(q \mid D)$ is the probability that an individual query term q will be found in the document stage of the HMM, and $P(q \mid GE)$ is the probability that the query term will be found in the General English stage of the HMM, but not in the document stage. The coefficients a_1 and a_0 are used to weight the respective components of the equation.

3.2 Document-Likelihood Models

The premise of the document-likelihood approach is that a language model can be generated for the query, and documents within the collection can be ranked by their probability of having been generated by the query's language model. In practice, this approach is most often used to enable expansion-based feedback as the terms in a query are generally too sparse to produce a reliable language model. [Zhai and Lafferty 2001] first introduce the idea of a query language model, and propose two methodologies for query language model construction: 1) a generative model of feedback documents and 2) a model that minimizes divergence over feedback documents. The generative model is a mixture model that generates a feedback document by mixing the query terms with a collection language model, which is taken to be a reasonable model of irrelevant content in a document. A document can then be generated from the resulting language model if it contains either the query terms or the collection language model. Conceptually, the divergence minimization language model estimates the query model by minimizing the average divergence between the query terms and the feedback documents. The estimated resulting query model is close to each feedback document model; however, in order to minimize the effect of general terms that may be common

to all the feedback documents, a regularization term is added to prefer documents that have a greater divergence from the collection language model.

The relevance model proposed by [Lavrenko and Croft 2001] can also be thought of as a document-likelihood model. Conceptually, the relevance model is a description of a user's information need which manifests itself in the form of a query. Given a collection of documents and a user's query, there exists a set of documents that are relevant to that query in the user's judgment. The ideal relevance model for a given query run on a specified document collection would be constructed from only the set of relevant documents within the collection; the relevance model in this framework is assumed to be a language model to which word probabilities are assigned. Each document relevant to the user's query then simply becomes a sample from the underlying relevance model.

The problem with this scenario is that in a typical retrieval environment we do not know the full set of relevant documents to a query and furthermore, we may not have any examples of documents which are relevant to the query. Lavrenko and Croft [2001] suggest a methodology that constructs a relevance model from a set of top ranked documents returned from a query. A relevance model is formally defined as the probability of observing a word w in a set of relevant documents R, or $p(w \mid R)$. The query q is also treated as a sample from R, although the sampling process that produces q is not necessarily the same as the process that generates w. Lavrenko and Croft formally derive a process whereby $p(w \mid R)$ can be estimated via $p(w \mid M_D)$, where M_D is the document model for a limited set of top-ranked documents returned from the query. They describe $p(w \mid M_D)$ as follows:

$$p(w \mid M_D) = \lambda \frac{tf(w,D)}{\Sigma_v tf(v,D)} + (1-\lambda)P(w \mid G) \tag{11}$$

Here $tf(w,D)$ is the number of occurrences of w in D, $\Sigma_v tf(v,D)$ is the total number of occurrences of all terms v in D, and $P(w \mid G)$ is the collection frequency of w divided by the total number of terms in the collection. The smoothing parameter λ is set empirically. This approach is elegant in that it can easily incorporate common information retrieval procedures which would not otherwise fit cleanly into a language modeling framework such as pseudo-relevance feedback or true relevance feedback. A linear interpolation method is used to smooth the MLE document models with the background model of English; smoothing parameters are set experimentally.

Subsequent work in relevance models by [Li 2005] treats a query as a short, special document and includes it in the documents that are used to approximate the relevance model to improve the robustness of the relevance modeling approach. In addition, instead of using a uniform prior as in the original relevance model, documents are assigned different priors based on their lengths and the probability of a term in the language model is adjusted by its probability in the language model of the corpus. This variant of the relevance modeling approach was applied to both a pseudo relevance feedback and true relevance feedback environment, and compared with the [Ponte and Croft 1998] query-likelihood approach as well as against the original [Lavrenko and Croft 2001] relevance model.

In later works, [Lavrenko et al. 2002] extend the relevance modeling approach to

Cross-Language Information Retrieval (CLIR). The proposed method "constructs an accurate relevance model in the target language, and uses that model to rank the documents in the collection [Lavrenko et al. 2002]." Their approach discusses two estimation strategies: one that assumes a parallel corpus (e.g., documents discussing the same topic in both query and target languages), and the other that assumes the existence of bilingual lexicon. In the former case, a joint probability model of word observations across the two languages is constructed; that is, a probability model is constructed that describes co-occurrence of words across languages that exist in documents related to the same topic. In the latter case, it is the lexicon which provides translation probabilities between words across the two languages. For cross-language information retrieval, the authors diverged from their earlier approach of ranking retrieved documents by the Probability Ranking Principle; instead they use Kullback-Leibler divergence [Kullback and Leibler 1951] as they found that this is "a more stable metric."

3.3 Model Comparison Approaches

The model comparison approach is first introduced by [Lafferty and Zhai 2001]. In their risk minimization 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 [Lafferty and Zhai 2001]." Within this context, a query is viewed as the output of a probabilistic process associated with a user U and a document is viewed as the output of a probabilistic process associated with an author or document source S. A user first selects an internal model Φ_Q having a probability distribution $p(\Phi_Q \mid U)$. A particular query q is then generated based on the parameters of that internal model with a probability of $p(q \mid \Phi_Q)$. Similarly, a document source selects an internal model Φ_D according to probability $p(\Phi_D \mid S)$, and then the probability of a generation of a particular document is given by the probability $p(d \mid \Phi_D)$.

In the Bayesian decision framework, there is an expected risk associated with every action of a given system. In this context, the particular action that the system performs is returning a document d_i in response to a query. The risk function for an action a can be modeled by understanding the loss $L(a)$ associated with that function. The function $R(d_i,q)$ which describes the risk associated with returning a particular document d_i, in response to a query is shown in the equation below:

$$R(d_i,q) = \sum_{R \in \{0,1\}} \int_{\phi_Q} \int_{\phi_D} L(\phi_Q,\phi_D,R) \times p(\phi_Q \mid q,U) p(\phi_D \mid d_i,S) p(R \mid \phi_Q,\phi_D) d\phi_D d\phi_Q \quad (12)$$

This is the basic retrieval formula based on risk minimization proposed by Lafferty and Zhai, which is used to calculate the ranking of documents d_i, returned in response to a query q. Lafferty and Zhai show how this risk minimization framework can be used to derive the "special cases" of the classical probabilistic model using a relevance-based loss function, and the query-likelihood language modeling approach using a distance-based risk function. The Kullback-Leibler (KL) divergence model, which is later elaborated in [Zhai and Lafferty 2001], is presented as a special case of the more general risk-minimization framework. In the KL-divergence model, the relevance

value of a document with respect to a query is measured by the probabilistic Kullback-Leibler divergence between the query model and document model. The problem of matching a query to a document thus reduces to a similarity or "distance" comparison which is similar to the classic vector-space model.

To address the sparse data problem, the authors explore an approach to query and document language model expansion based on Markov chains which is motivated by statistical translation methods of [Berger and Lafferty 1999]. The intent of the Markov chain is to model the user's browsing process, and the chain proposes a random walk alternating between queries and documents. Conceptually, the user is "surfing" through the word index for a given document collection, viewing the documents which contain that word, and then refining their information need as they go along. Practically, this approach calculates "the posterior probabilities of words according to the translation model for generating the query and a prior distribution on initial terms selected by the user [Lafferty and Zhai 2001]." Mathematically, this approach can be represented by the following formula describing the probability of occurrence for a word w in a query q:

$$p(w \mid \phi_q) \propto \sum_i t(q_i \mid w) p(w \mid U) \tag{13}$$

Here, $p(w \mid \Phi_q)$ is an estimate for the language model, $t(q_i \mid w)$ is the translation model which describes the likelihood of translation between query term q_i and word w, and $p(w \mid U)$ is the likelihood that the user will start their "surfing" from the initial word w. The analogous formula for document expansion is shown below:

$$p(w \mid \phi_d) \propto t(d \mid w) p(w \mid U) \tag{14}$$

3.4 Statistical Translation Models

In their statistical translation model, [Berger and Lafferty 1999] propose that the formulation of a query is really the distillation of a user's information need into a succinct form. The distillation from a "fat" document to a "skeletal" query, the authors propose, "is a form of translation from one language to another [Berger and Lafferty 1999]." Berger and Lafferty represent this document to query translation process using a statistical model. They characterize the translation process as having two stages. In the first stage, the translation analyst chooses a word w from the document according to a distribution $l(w \mid d)$ which is called the document language model. In the second stage, that word w is translated into a word or phrase q using a translation model $t(q \mid w)$. Thus, the statistical translation model for a single query term q can be described by the following equation:

$$p(q \mid d) = \sum_{w \in d} l(w \mid d) t(q \mid w) \tag{15}$$

This model must be applied n times to account for a query containing n terms; the number n in this model is chosen according to a sample size model $\Phi(n \mid d)$. Berger and Lafferty propose that a Poisson distribution with mean $\lambda(d)$ can be used to calculate the sample size model.

$$\phi(n \mid d) = e^{-\lambda(d)} \frac{\lambda(d)^n}{n!} \tag{16}$$

Applying this assumption leads to the complete statistical translation model, which Berger and Lafferty call Model 1:

$$p(q = q_1, q_2, \ldots, q_n \mid d) = e^{-\lambda(d)} \prod_i e^{\lambda(d) p(qi \mid d)} - 1 \tag{17}$$

It should be noted that if each word can be translated only to itself (e.g., $p(q \mid t)=1$ only when $q=t$) this model decomposes into the query-likelihood model proposed by Ponte and Croft [1998]. Berger and Lafferty call this simplest version of the translation model Model 0.

The parameters for the translation model were set empirically using the Expectation Maximization (EM) algorithm [Dempster et al. 1977]. A simple linear mixture model is used to combine the background unigram model for the corpus and the EM-trained translation model. The smoothing parameters were derived empirically by optimizing the algorithm on the TREC Spoken Document Retrieval data. While significant improvements over a baseline *tf-idf* approach were reported by this method, the experiments required a large corpus of training data which may not be practical in all cases. An interesting corollary to this approach is that it is naturally extensible to the problem of multi-lingual information retrieval.

[Murdock and Croft 2004] extend the translation model approach to sentence retrieval. The motivation for this application of translation models stems from a Question Answering (QA) application; in general, most QA systems use a passage retrieval system for the first stage of processing. The more accurately a passage retrieval system can retrieve succinct segments of text, the better the QA system is likely to perform overall. A high-quality sentence retrieval system would be a strong candidate for application to the QA domain. However, sentences, which are much smaller, than documents are too short to accurately estimate a language model. Murdock and Croft approach this problem by using a translation model to judge similarity between a words in query and words in a candidate sentence. The translation model allows for a looser matching strategy which identifies the relationships between corresponding terms which mean the same thing or are related to one another, but which are not the same term. Murdock and Croft make use of the IBM translation Model 1 [Brown et al. 1990] to rank documents according to their translation probability, given the query. IBM Model 1 assumes that all alignments are possible between the source sentence and target sentence, and constructs synthetic training data in the absence of the availability of a training corpus of queries and relevant documents.

In related research, Jin and use language models for document title generation. Rather than using the document directly as the knowledge source for the document title generation, they introduce the idea of a "distilled information source" which is a sample of important content words from the original document. The optimal title for the document can then be generated from this distillation. The underlying premise of this method is that title generation is a "reverse" information retrieval task – in other

words, the perfect title for a document would be the same as the query for which that document is an "ideal" response. In an information retrieval paradigm, the generated titles for each document can be evaluated against the queries that are presented to the system and the documents ranked accordingly. This approach is novel in that it provides a bridge from language modeling and information retrieval to the related tasks of text summarization and categorization.

3.5 Cluster-based Language Models

In the general sense, cluster-based language modeling approaches use document clustering to organize collections around topics. Each cluster is then assumed to be representative of a topic, and a language model can be created for the cluster. Cluster-based language models are most commonly used in the context of Topic Detection and Tracking, but have been incorporated into other IR frameworks [Liu and Croft 2004] as well. Although we choose to break out cluster-based language models as a separate section in this paper, one could just as easily group individual cluster-based modeling approaches with query-likelihood, document-likelihood or model comparison approaches, depending on the specific nature of the language model employed to describe the clusters. (We are not aware of any cluster-based language modeling approaches to date which have employed statistical translation methods.) In some sense, one might even say that cluster-based language modeling can trace its roots to [Kalt 1996], which focused on statistical models for topics as opposed to individual documents.

One of the earliest cluster-based language modeling approaches was employed in the context of distributed information retrieval. [Liu and Croft 2004] propose that "the task of a distributed retrieval system is first to determine which topics are best for a query and then to direct the searching process to those collections containing the topics." They consider that the problem of determining the most suitable topics for a given query can be addressed using a generative model; that is, the best topics for a given query are those which have a language model that is most likely to generate the query. Collection selection for distributed IR is then performed by selecting the collections which contain the best topics.

Later work by [Liu and Croft 2004] proposes the use of cluster-based language models in the context of an ad-hoc retrieval framework. Liu and Croft propose two language models for cluster-based retrieval: the first is used in ranking and retrieving clusters and the other uses cluster language models to smooth individual document language models within the cluster. These cluster-based models are then integrated into both a query-likelihood and relevance model framework. Empirical results show that cluster-based retrieval can potentially be more effective than document based retrieval. It is specifically interesting to note the success of the cluster-based smoothing methods, and that "clusters generated by static clustering tend to produce better-quality cluster models for smoothing purposes than those generated by query-specific clustering [Liu and Croft 2004]." Static clustering, performed without query text, looks at all documents in the collection and generates clusters that provide better coverage for all aspects of a given topic. The authors believe that the static clustering methods outperform query-based clustering methods as query-based clusters may

contain an inherent bias to a specific interpretation of a query term.

In a more recent approach, Kurland [Kurland et al. 2005] proposes that a document retrieved as a part of pseudo-relevance feedback may be considered as a "rendition" of the original query. Documents which are good renditions of the query may be considered to be pseudo-queries and considered to be a wholesale replacement for the original query itself. Kurland's approach proposes that once we have created an initial set of pseudo-queries, the process can be repeated so that in the next iteration the algorithm is searching for the documents that are the best rendition of the pseudo-queries. After the process is complete, the result should be a set of distilled pseudo-queries which are optimally informative of the user's information need. Kurland proposes three algorithms (the Viterbi Doc-Audition algorithm, the Doc-Audition algorithm, and the Cluster-Audition algorithm) to score documents as candidate renditions of the pseudo-query.

- The motivation behind this approach is to increase the "aspect recall" of the system. The problem of "aspect recall" is described in [Buckley 2004; Harman and Buckley 2004] and can be categorized in one of four ways:
- The IR system emphasizes one aspect of a query, and misses other required terms in the query
- The IR system emphasizes one aspect of a query, but misses other aspects
- The IR system fails to properly combine aspects in a query when returning query results
- The IR system emphasizes an irrelevant aspect

Query drift is an issue that this particular algorithm is particularly prone to. "Query drift" [Mitra et al. 1998] is a problem encountered by many automatic query expansion approaches; when expanding a query if non-relevant terms are included as a part of the expansion the resulting query may "drift" from the original information need. Kurland addresses this problem by using re-scoring techniques to periodically re-align the pseudo-queries with the original queries. The empirical results obtained from this methodology are promising when compared against the baseline relevance modeling approach [Song and Croft 1999].

Other cluster-based language modeling approaches include Zhang [Zhang et al. 2005], who models the generation of clusters using a Dirichlet process mixture model, where the base distribution can be treated as the prior of general English model and the precision parameter which controls the random generation process for creating new clusters.

4. SMOOTHING STUDIES

Each of the approaches discussed in this paper makes use of smoothing techniques. Smoothing is an integral part of the language modeling paradigm and the performance of the smoothing component of a language model is essential to the overall performance of the model. Several classes of smoothing strategies have been proposed; the most common has been described as parameter smoothing [Liu and Croft 2005], and uses linear interpolation to influence the roles that multiple knowledge sources play in smoothing out the probability distributions of a language model. [Zhai and Lafferty 2004] studied three approaches to smoothing: Jelinek-Mercer smoothing, Dirichlet

priors and absolute discounting, as well as the backoff versions of these methods. Five test collections were used to examine the effects of each of these smoothing mechanisms. For title queries (short queries), there was a clear ordering among the methods in terms of precision results; Dirichlet priors performed better than absolute discounting, which performed better than Jelinek-Mercer. Overall, Dirichlet priors had an average precision performance that was significantly better than the other two methods. For longer queries, both Jelinek-Mercer and Dirichlet priors have a better performance than absolute discounting.

A later study by [Smucker and Allan 2007] investigate the causes behind the improved performance yielded by Dirichlet prior smoothing over Jelinek-Mercer smoothing for short queries. Short queries are especially important as they are to be most characteristic of a typical query that would be posed by a user. Both Dirichlet prior and Jelinek-Mercer linearly combine the maximum likelihood estimated (MLE) document model with the MLE model of the collection. Both are discounting smoothing methods that reduce the probability of the words seen in the document and reallocate the probability mass to words not seen in the document. The only difference between the two smoothing methods that can be observed is that Dirichlet prior smooths longer documents less and Jelinek-Mercer smooths all documents to the same degree. Intuitively this makes sense since it stands to reason that the MLE model of a short document contains less observed information than the MLE model of a long document. Smucker and Allan found that Dirichlet prior's performance advantage comes more from its penalization of shorter documents rather than from its estimation process for long documents; in other words, Dirichlet prior is able to correctly reduce the scores of short documents. This research points to the importance of factors such as document length in the language modeling process.

5. SEMANTICS AND THE LANGUAGE MODELING FRAMEWORK

Semantic smoothing [Liu and Croft 2005] is one area in which there are significant opportunities for the integration of context-sensitive semantic knowledge into the language modeling framework. Translation models may be thought of as one form of semantic smoothing, since a translation model provides a mechanism for mapping based on document-query training sets. However, translation models do not provide any mechanism for sense disambiguation; as such, terms that appear with high likelihoods in two different senses will be mixed together. For example, the term "apple" may appear in conjunction with the word "computer" and the word "pie" both with high likelihoods.

[Zhou et al. 2006] proposes a context-sensitive semantic smoothing as a part of the language modeling framework, which introduces the concept of a topic signature for a document or query and then uses these topic signatures to train the translation model. Zhou defines a topic signature to be an order-free relationship between two concepts, where a concept represents a set of synonymous terms within a domain. For the biomedical domain, the Unified Medical Language System (UMLS) Metathesaurus (http://www.nlm.nih.gov/research/umls) is one source for concept definitions. Incorporation of semantic information by way of topic signatures alleviates the sense disambiguation (synonymy and polysemy) problems that are incurred with previous

translation model approaches. This approach presents a different view of a document representation within the language modeling framework; a document is presented as a weighted set of topic signatures and concepts. An Expectation-Maximization (EM) [Dempster et al. 1977] based training method is used to train the context-sensitive model. The empirical results of this semantic approach showed significant improvements over the baseline on TREC Genomics 2004 [Hersch 2004] and 2005 [Hersch 2005] data. One drawback to this ontology-based approach, however, is that it requires the existence of domain-specific ontological resources which are not often available. This may make such an approach difficult to apply to a broad range of fields.

Another approach that may lend itself to incorporation within a semantic smoothing framework is Probabilistic Latent Semantic Analysis (PLSA) [Hofmann 1999]. The foundation of the PLSA approach is an underlying Aspect Model which proposes that we can define words and documents in terms of "aspects" which are associated with a latent class variable. The Aspect Model has several intuitively appealing features. First, by conditioning words and documents on a latent variable, the zero-frequency problem is addressed. Secondly, a priori knowledge is not required about the concepts within the corpus for the algorithm to work effectively. And finally, the usage of probabilistic methods defines a generative model of the data which is better able to address common text processing issues such as synonymy and polysemy. Recent papers by [Banerjee and Han 2008; Banerjee and Han 2009a; Banerjee and Han 2009c] explore the incorporation of PLSA into the language modeling framework. Specifically, they use PLSA to derive an Aspect-Based Relevance Language Model, which is then used to model a semantic Question Context.

6. LANGUAGE MODELING AND XML RETRIEVAL

[Ogilvie and Callan 2003] believe that the information contained in the structure of an XML document can be used to improve document retrieval for structured knowledge sources. In order to leverage this information, they model document structure as an integral part of a document language model. Specifically, XML documents are modeled as trees where each node in the tree correspond directly with tags present in the document. For each document node in the tree, a language model can then be estimated. Language models for leaf nodes with no children are estimated directly from the text of the node. The language models for other nodes in the tree are estimated by taking a linear interpolation the language model of the text contained in that node alone with the language models of all child nodes. The language models for each node are then smoothed via an interpolation with a collection language model. The collection model used for the interpolation may be specific to the node type, which provides some measure of context sensitive smoothing, or the collection model may be one large model estimated from everything in the corpus, which gives a larger sample size.

In a subsequent work, [Ogilvie and Callan 2006] focus on the problem of parameter estimation for the hierarchical language model for XML documents. Specifically, the parameters that are considered are the interpolation parameters which are used to combine the language models at a given node with the language models of the node's children. Unlike their previous work [Ogilvie and Callan 2003], Ogilvie and Callan

make a simplifying assumption; they do not recursively smooth the language models when they traverse the structure of the XML document. Instead, they linearly interpolate the parent's unsmoothed language model, each child's unsmoothed language model, the document's unsmoothed language model, and the collection language model. This simplification allows for a cleaner formulation of the parameter estimation problem which enables application of a modified version of the EM algorithm which the authors call Generalized Expectation Maximization (GEM). The proposed approach first trains the EM algorithm, but observes that using positive examples alone places the most weight on the document language model which results in very poor retrieval performance. To counter this effect, negative examples are included in the model; negative examples are non-relevant components that come from documents that contain relevant components. Conceptually, the GEM algorithm maximizes the probability that the language models of the positive samples will generate the query term while minimizing the likelihood that the language models of the negative samples will not generate the query term. The approaches presented in [Ogilvie and Callan 2003; Ogilvie and Callan 2006] are validated empirically on data from the INEX CO tasks, with performance that is mixed but shows promise for further development.

7. FUTURE TRENDS

[Allan et al. 2003] presents a comprehensive discussion of near and longer term challenges in Information Retrieval to which language modeling may be applied. These challenge areas include "retrieval models, cross-lingual retrieval, Web-search, user modeling, filtering, Topic Detection and Tracking (TDT), classification, summarization, question answering, meta-search and distributed retrieval, and multimedia retrieval, information extraction and testbeds." Although this paper was written several years ago, the majority of the challenges and issues presented are still pertinent to state-of-the-art Information Retrieval systems today. It is beyond the scope of this paper to discuss each of the challenge areas in detail; instead we will focus on three areas that are of particular interest to us: retrieval models, cross-language information retrieval and question answering.

Empirically, retrieval models that incorporate relatively simple language modeling techniques have produced promising results over the past decade. The majority of these approaches have used a unigram language model, although a few have explored approaches based on bi-gram and tri-gram models. This relatively simplistic view of the document as a "bag of words" is an opportunity for future research. In his survey of statistical language modeling techniques, [Rosenfeld 2000] says, "Ironically, the most successful statistical language modeling techniques use very little knowledge of what language really is ... only a handful of attempts have been made to date to incorporate linguistic structure, theories or knowledge." The incorporation of these types of formalized structures into a language modeling framework may help to increase the effectiveness of language modeling approaches. In addition, further improvements in the language modeling paradigm "are likely to require a broad range of techniques in addition to language modeling [Allan et al. 2003]." This is especially true if one considers the changing landscape of information resources; grass-roots innovations such as wikis, blogs, social bookmarking and even social networking

applications such as MySpace and Facebook present unique challenges and opportunities for modeling language. Hierarchical models such as those proposed by Ogilvie [Ogilvie and Callan 2003] are promising when one considers retrieval applications that merge unstructured and semi-structured data; it may be likely that future approaches build on this foundation when considering language data from diverse sources.

A second challenge area is cross-language information retrieval research. [Allan et al. 2003] states that "though initially the Web was dominated by English speakers, now less than half of existing web pages are in English." If this statement was true in 2002, today the importance of cross-language information retrieval is even more evident. From a humanitarian standpoint, one may consider that we have a social responsibility to ensure that cross-language information search and retrieval techniques are not limited to those languages for which large amounts of data are available – such as English, Spanish, French, Arabic, and Chinese. In order to effectively extend language modeling techniques to languages for which less data is available, existing methods should look to develop techniques for which little or no training data is required. If such research is not undertaken, the distinction between those who have access to information and those who do not will be made along language boundaries, and the information gap will widen as time goes on.

A final challenge area for language modeling that is of particular interest is Question Answering (QA). Recent developments in Question Answering research have started to address many of the topics that were discussed in [Allan et al. 2003]. Recent TREC conferences [Voorhees 2006; Voorhees 2005a; Voorhees 2005b; Voorhees and Harman 2005] have considered reliable factoid QA, interactive QA, development of user and session models which require that a given question be cognizant of the QA dialogue that preceded it and novelty detection within a QA setting (e.g. "Tell me something interesting about the topic that I have not seen yet."). In addition, recent QA approaches [Hovy et al. 2002] have considered merging of structured and semi-structured information (such as information from knowledge sources such as Wikipedia and WordNet) to improve the QA reliability. The TREC Genomics Track [Hersh et al. 2006] can also be thought of as a QA exercise; in those experiments, researchers must return the answers to broad questions with some source document context. Despite the recent focus in this area, however, there are still many aspects of QA that require development or improvement. Within the QA application area, language modeling has largely been used to improve the first stage of processing which collects candidate documents and passages that are then passed on to an Information Extraction (IE) engine which determines the specific response words or phrases. A future goal for language modeling may be to provide for extensions that allow for integration with the second stage of QA – either by the incorporation of knowledge patterns or via statistical mechanisms that can be integrated into the IE or Natural Language Processing (NLP) techniques commonly used in downstream QA processing.

One QA research area which is promising for the application of language modeling techniques is Answer Validation. In recent years, Answer Validation has become a topic of significant interest within the QA community. In the general sense, one can describe Answer Validation as the process that decides whether a Question is

correctly answered by an Answer according to a given segment of supporting Text. In the seminal work on Answer Validation, [Magnini et al. 2002] presents an approach that uses redundant information sources on the Web; they propose that the number of Web documents in which the question and the answer co-occurred can serve as an indicator of answer validity. [Banerjee and Han 2009] propose the use of language modeling methodologies for Answer Validation, using corpus-based methods that do not require the use of external sources. Specifically, they propose the development of an Answer Credibility score which quantifies reliability of a source document that contains a candidate answer. To insert this model into the Answer Validation process, they propose an interpolation technique that modulates the answer score during the process using Answer Credibility.

8. CONCLUSION

Statistical language modeling is a technique that has been in use in natural language processing and speech recognition for several decades. The recent application of statistical language modeling to information retrieval has proven to be immensely successful, and yielded empirical results that have surpassed earlier retrieval engines. Smoothing is an integral part of the language modeling framework and no language modeling approach can be discussed without discussing the smoothing approaches that are taken in conjunction. There are, however, many challenges ahead if statistical language modeling is to become an integral part of related domains such as cross-lingual information retrieval, question answering and user-context modeling. We expect significant research to be done in these areas in the years ahead which will, hopefully, contribute to a globally-integrated information environment.

REFERENCES

ALLAN, J., J. ASLAM, N. BELKIN, C. BUCKLEY, J. CALLAN, W. B. CROFT, S. DUMAIS, N. FUHR, D. HARMAN, D. HARPER, D. HIEMSTRA, T. HOFMANN, E. HOVY, W. KRAAIJ, J. LAFFERTY, V. LAVRENKO, D. LEWIS, L. LIDDY, R. MANMATHA, A. MCCALLUM, J. PONTE, J. PRAGER, D. RADEV, P. RESNIK, S. ROBERTSON, R. ROSENFELD, R. ROUKOS, M. SANDERSON, R. M. SCHWARTZ, A. SINGHAL, A. SMEATON, H. TURTLE, E. VOORHEES, R. WEISCHEDEL, J. XU, AND C. ZHAI. 2003. Challenges in information retrieval and language modeling. *Report of a workshop held at the center for intelligent information retrieval.* 37 (September):31–47.

BAHL, L. R., P. F., BROWN, P. V. DE SOUZA, AND R. L. MERCER. 1989. A tree-based statistical language model for natural language speech recognition. *Communications of the ACM* 37:1001–1008.

BANERJEE, P. AND H. HAN. 2008. Incorporation of Corpus-Specific Semantic Information into Question Answering Context. In *CIKM 2008 – Ontologies and Information Systems for the Semantic Web Workshop Napa Valley*, USA.

BANERJEE, P. AND H. HAN. 2009a. Modeling Semantic Question Context for Question Answering. To appear in FLAIRS 2009.

BANERJEE, P. AND H. HAN. 2009b. Answer Credibility: A Language Modeling Approach to Answer Validation. To appear in NAACL-HLT 2009.

BANERJEE, P. AND H. HAN. 2009c. From Question Context to Answer Credibility: Modeling Semantic Structures for Question Answering Using Statistical Methods. To appear in IKE 2009.

BERGER, A. AND J. LAFFERTY. 1999. Information retrieval as statistical translation. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in*

information retrieval, 222–229.

BROWN, P., S. DELLA PIETRA, V. DELLA PIETRA, F. JELINEK, J. LAFFERTY, R. MERCER, AND P. ROOSSIN. 1990. A statistical approach to machine translation. *Computational Linguistics*. 16:79–85.

BUCKLEY, C. 2004. Why current IR engines fail. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval Sheffield*. UK, 584–585.

DEMPSTER, A. P., LAIRD, N. M. AND D. B. RUBIN. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

HARMAN, D. AND C. BUCKLEY. 2004. The NRRC reliable information access (RIA) workshop. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. Sheffield, UK, 528–529.

HERSCH, W. 2004. TREC 2004 Genomics Track Overview. In *On-line Proceedings of the Thirteenth Text Retrieval Conference*.

HERSCH, W. 2005. TREC 2005 Genomics Track Overview. In *On-line Proceedings of the TREC 2005 Genomics Track Overview*.

HERSH, W., A. COHEN, P. ROBERTS, AND H. K. REKAPALLI. 2006. TREC 2006 Genomics Track Overview. In *Online proceedings of the 2006 Text Retrieval Conference*.

HIEMSTRA, D. 1999. A linguistically motivated probabilistic model of information retrieval. In *Research and Advanced Technology for Digital Libraries – Second European Conference*. ECDL'98, 569–584.

HOFMANN, T. 1999. Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development In Information Retrieval*.

HOVY, E., U. HERMJAKOB, AND C. Y. LIN. 2002. The Use of External Knowledge in Factoid QA. In *NIST Special Publication*. Gaithersburg, Maryland, 644–652.

JELINEK, F. AND R. L. MERCER. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice Amsterdam*. Netherlands.

JELINEK, F., R. L. MERCER, L. R. BAHL, AND J. K. BAKER. 1977. Perplexity – a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*. 62:S63.

JIN, R. AND A. HAUPTMANN. 2001. Learning to Select Good Title Words: An New Approach based on Reversed Information Retrieval. *Proceedings of the Eighteenth International Conference on Machine Learning*. 242–249.

JURAFSKY, D. AND J. H. MARTIN. 2000. Speech and language processing: Prentice Hall Upper Saddle River, NJ.

KALT, T. 1996. A New Probabilistic Model of Text Classification and Retrieval. *Technical Report*. University of Massachusetts, Amherst, Massachusetts.

KATZ, S. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics Speech and Signal Processing*, 35:400–401.

KRAAIJ, W. AND M. SPITTERS. 2003. Language Models for Topic Tracking. In Language Models for Information Retrieval, W. B. Croft and J. Lafferty, Eds.: Kluwer Academic Publishers.

KULLBACK, S. AND R. A. LEIBLER. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*. 22:79–86.

KURLAND, O., L. LEE, AND C. DOMSHLAK. 2005. Better than the real thing?: iterative pseudo-query processing using cluster-based language models. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. Salvador, Brazil, 19–26.

LAFFERTY, J. AND C. ZHAI. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval New Orleans*. Louisiana: ACM Press, 111–119.

LAVRENKO, V. AND W. B. CROFT. 2001. Relevance based language models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.* 120–127.

LAVRENKO, V., M. CHOQUETTE, AND W. B. CROFT. 2002. Cross-Lingual Relevance Models. In *Proceedings of the 25th annual international ACM SIGIR Tampere.* Finland, 175–182.

LI, X. 2005. Improving the Robustness of Relevance Based Language Models. *CIIR Technical Report, University of Massachusetts.* Amherst.

LIU, X. AND W. B. CROFT. 2004. Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* Sheffield, UK, 186–193.

LIU, X. AND W. B. CROFT. 2005. Statistical Language Modeling For Information Retrieval. *Annual Review of Information Science and Technology.*

MAGNINI, B., M. NEGRI, R. PREVETE, AND H. TANEV. 2002. Is It the Right Answer? Exploiting Web Redundancy for Answer Validation. In *Association for Computa-tional Lingustistics (ACL) 2002.* Philadelphia, PA, 425–432.

MANNING, C. D. AND H. SCHÜTZE. 1999. Foundations of Statistical Natural Language Processing: The MIT Press.

MANNING, C. D., P. RAGHAVAN AND H. SCHUTZE. 2007. Introduction to Information Retrieval: Cambridge University Press.

MILLER, D. R. H., LEEK, T. AND R. M. SCHWARTZ. 1999. A hidden Markov model information retrieval system. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* 214–221.

MITRA, M., A. SINGHAL, AND C. BUCKLEY. 1998. Improving automatic query expansion. In *Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Melbourne, Australlia, 206–214.

MURDOCK, V. AND W. B. CROFT. 2004. Simple translation models for sentence retrieval in factoid question answering. In *ACM SIGIR Workshop on Information Retrieval for Question Answering.*

OGILVIE, P. AND J. CALLAN. 2003. Using Language Models for Flat Text Queries in XML Retrieval. In *INEX 2003 Workshop Proceedings.*

OGILVIE, P. AND J. CALLAN. 2006. Parameter Estimation for a Simple Hierarchical Generative Model for XML Retrieval. In *Advances in XML Information Retrieval and Evaluation.* vol. 3977: Springer, 211.

PONTE, J. M. AND W. B. CROFT. 1998. A language modeling approach to information retrieval. In *21st annual international ACM SIGIR conference on Research and development in information retrieval Melbourne.* Australia, 275–281.

ROBERTSON, S. AND K. JONES. 1997. Simple proven approaches to text retrieval. *Cambridge University Computer Laboratory Technical Report.*

ROSENFELD, R. 2000. Two decades of statistical language modeling: where do we go from here?. *Proceedings of the IEEE,* 88:1270–1278.

SALTON, G. AND M. J. MCGILL. 1986. Introduction to Modern Information Retrieval: McGraw-Hill, Inc. New York, NY, USA.

SHANNON, C. E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 30:50–64.

SMUCKER, M. AND J. ALLAN. 2007. An Investigation of Dirichlet Prior Smoothing's Performance Advantage. *Technical Report, University of Massachusetts.* Amherst, Amherst, Massachusetts.

SONG, F. AND W. B. CROFT. 1999. A general language model for information retrieval. *Proceedings of the eighth international conference on Information and knowledge management.* 316–321.

VOORHEES, E. M. 2005a. Overview of the TREC 2005 Question Answering Track. In *Online proceedings of the 2005 Text Retrieval Conference.*

VOORHEES, E. M. 2005b. Overview of the TREC 2005 Robust Retrieval Track. *On-line Proceedings of the Thirteenth Text Retrieval Conference.*

VOORHEES, E. M. 2006. Overview of the TREC 2006 Question Answering Track. In *Online*

proceedings of 2006 Text Retrieval Conference.

VOORHEES, E. M. AND D. HARMAN. 1999. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Online proceedings of 1999 Text Retrieval Conference.*

VOORHEES, E. M. AND D. K. HARMAN. 2005. TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing): The MIT Press.

WITTEN, I. H. AND T. C. BELL. 1991. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory.* 37:1085–1094.

XU, J. AND W. B. CROFT. 1999. Cluster-based language models for distributed retrieval. *Proceedings of the 22nd annual international ACM SIGIR.* 254–261.

ZHAI, C. AND J. LAFFERTY. 2001. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the tenth international conference on Information and knowledge management Altanta.* GA: ACM Press New York, NY, USA, 403–410.

ZHAI, C. AND J. LAFFERTY. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS),* 22:179–214.

ZHANG, J., Z. GHAHRAMANI, AND Y. YANG. 2005. A probabilistic model for online document clustering with application to novelty detection. In *Advances in Neural Information Processing Systems 17.* vol. 17, Y. W. Lawrence K. Saul, Leon Bottou, Ed.: MIT Press, 1617–1624.

ZHOU, X., X. HU, X. ZHANG, X. LIN, AND I. Y. SONG. 2006. Context-sensitive semantic smoothing for the language modeling approach to genomic IR. In *29th annual international ACM SIGIR Seattle.* WA, USA, 170–177.

ZIPF, G. K. 1949. Human behavior and the principle of least effort: Addison-Wesley Press Cambridge, Mass.

Protima Banerjee is currently a doctoral candidate in the Information Science and Technology School at Drexel University. She received her undergraduate degree in Electrical Engineering from Cornell University in 1995 and her Masters degree in Computer Science from Rensselaer Polytechnic Institute in 1997. Ms. Banerjee has held technical and leadership positions designing and developing information systems for Oracle Corporation, Actium Technologies and Idea Integration. Since 2001, she has been a Systems Engineer with the Lockheed Martin Corporation. Ms. Banerjee's research interests include information retrieval, question answering and semantic information integration.

Hyoil Han is an assistant professor at Drexel University. She obtained a B.S. degree from Korea University and an M.S. degree from Korea Advanced Institute of Science and Technology (KAIST). She then worked for the Samsung Electronics and Telecommunication Network Lab at Korea Telecom as a member of the technical staff before obtaining a Ph.D. in Computer Science and Engineering from the University of Texas at Arlington in 2002. Her research areas lie in the merging of techniques from the fields of databases (DB) and artificial intelligence (AI) and applying the new combined techniques to biomedical informatics and the Semantic Web with an emphasis on data/text mining, and data integration/management. She has published more than 35 papers in refereed literature.