

EXPLICITLY CONSIDERING RELEVANCE WITHIN THE LANGUAGE MODELING FRAMEWORK

Leif Azzopardi

Department of Computing Science, University of Glasgow, Scotland
leif@dcs.gla.ac.uk

Thomas Roelleke

Department of Computer Science, Queen Mary, University of London, England
thor@dcs.qmul.ac.uk

Keywords: Information Retrieval, Probabilistic Language Models, Relevance

Abstract: Whilst the event of relevance is central to the Binary Independence Retrieval model, Language Modeling focuses on the estimation of the document model. In this paper, we review the different past formulations of the Language Modeling (query likelihood) approach. We find that these previous formulations largely ignore relevance by making implicit or explicit assumptions. The main contribution of this work is an alternative formulation that specifically relates relevance and language modeling in a sound probabilistic framework. This leads to valuable insights into the application of Language Modeling to Information Retrieval, including how the approach handles relevance information and how the approach can be further developed.

1 Introduction

The adaptation of statistical language modeling techniques to *ad hoc* retrieval was proposed in 1998, and is typically referred to as the Language Modeling approach (Ponte and Croft, 1998). Since then a steady stream of research into Language Modeling has been generated (Miller et al., 1999; Hiemstra, 2001; Zhai and Lafferty, 2001b; Azzopardi, 2005), and it has become widely accepted as an effective and intuitive retrieval model. However, there have been some concerns raised about how relevance is actually dealt with in the model (Sparck-Jones et al., 2003; Robertson, 2005).

In traditional probabilistic retrieval models (such as the Binary Independence Model (Robertson and Sparck-Jones, 1976)), the question the system attempts to answer is, “How probable is the event of relevance given the document and the query?”. This probability is denoted as $P(r|d, q)$ which can not be estimated directly. And so Bayes theorem is employed such that the probability of the document given relevance and the query (i.e. $P(d|r, q)$) is estimated instead. Central to this estimate is the notion of relevance, where the document score is conditioned by relevance. This is important, because under this decomposition the Probability Ranking Principle (PRP) (Robertson and Sparck-Jones, 1977) is upheld,

which guarantees that the ranking is optimal.

In the Language Modeling approach, the system answers an entirely different question, “What is the probability that this query came from this document?”. This probability is denoted $P(q|d)$ and is often referred to as the query likelihood. Here, there is no explicit event of relevance within the model. It is therefore unclear how the PRP would be upheld. Consequently, the model requires the following assumption to be engaged: that the query likelihood is correlated with a document’s relevance. This assumption has led to some criticism of the model and has sparked a debate concerning the model’s theoretical underpinnings (Sparck-Jones et al., 2003; Robertson, 2005). It has been suggested that without the explicit event of relevance within the model, estimation of the model when relevance feedback information is available is unclear and problematic (Sparck-Jones et al., 2003)¹. This criticism has yet to be adequately addressed.

In this paper, we provide an alternative formulation of the language modeling approach that explicitly defines relevance within the estimation of the query

¹While there are a number of methods which can use feedback for query expansion or query term re-weighting, these methods do not update a model of relevance (Zhai and Lafferty, 2001a; Hiemstra, 2002; Miller et al., 1999).

likelihood. This alternative formulation leads to a better understanding of relevance within the LM approach, that not only that illuminates some of the limitations of using the query likelihood, but also introduces new directions in which to extend and further develop the Language Modeling approach in a theoretically sound manner. The remainder of this paper is structured as follows: In the next section, we provide an overview of the different formulations of Language Modeling for ad hoc retrieval. We explain what implicit or explicit assumptions each formulation makes regarding relevance and how this leads to ranking based upon the query likelihood. Section 3 discusses these assumptions and their shortcomings in addressing relevance. In section 4, we present a different formulation of the query likelihood that explicitly defines the event relevance within the language modeling framework. This alternative involves the application of the Theorem of Total Probability to incorporate relevance and leads to two different approaches to estimating the query likelihood. In Section 5, we consider various ways in which to estimate the query likelihood under different interpretations. In Section 6, we discuss how relevance feedback information could be incorporated within these different interpretations. Finally, we conclude the paper in Section 7 with a summary of our contribution and directions for further research.

2 Background

There are four main query likelihood formulations of the Language Modeling approach² for ad hoc information retrieval which have been motivated from different points of view. These were put forward by Ponte and Croft (1998), Miller et al. (1999), Hiemstra (2001) and Lafferty and Zhai (2003). Each formulation treats the notion of relevance within the language modeling framework differently, but essentially the ranking of documents is performed through the same estimated probability - the probability of the query being generated from the document, $P(q|d)$. In this section, we present a summary of each formulation which contextualizes our contribution and we shall use the following notation. Let d denote a document, q denote the query, and let relevance be defined

²There are many other types of Language models such as the relevance based language models (Lavrenko and Croft, 2001) or loss functions (Zhai and Lafferty, 2004) which provide a more complex view of relevance than simply treating relevance as binary random variable. Here, our focus is solely on language models that rank according to the probability of a query given a document i.e. $P(q|d)$.

as a binary random variable where r denotes the event of relevance and \bar{r} which denotes the event of non-relevance. The query q is assumed to be composed of a sequence of k query terms t_i , i.e. $q = \{t_1, \dots, t_k\}$. For document modeling, we adopt the popular multinomial model (Hiemstra, 2001; Miller et al., 1999), where the document is characterized by a distribution over the vocabulary, i.e. the probability of term given the document, $P(t_i|d)$.

2.1 Ponte and Croft's Formulation

Ponte and Croft (1998) adapt language modeling from statistical natural language processing and apply predictive text models to Information Retrieval. In this seminal work, they use the probability of a query given a document, $P(q|d)$ to approximate the probability of a document given the query and relevance, $P(d|q, r)$. They arrive at this conclusion by first assuming that the probability $P(d|q, r)$ can be approximated by the probability of a document given the query, $P(d|q)$ (as shown in Equation 1). Through the application of Bayes' theorem in Equation 2, they then obtain $P(q|d)$.

$$P(d|q, r) \approx P(d|q) \quad (1)$$

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \quad (2)$$

$$\propto P(q|d) \quad (3)$$

Here, the probability of the query, $P(q)$, is a constant, and the probability of the document, $P(d)$, is assumed to be constant for all documents. The final ranking is therefore proportional to the $P(q|d)$ as shown in Equation 3.

The pivotal assumption, **PC.A1** engaged is that the $P(q|d)$ is correlated with the probability of document being relevant $P(d|q, r)$ (Ponte and Croft, 1998). The intuition behind the assumption is that a document is more likely to be relevant if that document is more likely to produce the query. This is a very appealing argument, which Ponte and Croft (1998) claim makes the Language Modeling approach an explanatory model of retrieval. However, by engaging this assumption, they largely side step the problems associated with modeling relevance, and instead rely upon the intuition and the demonstrated retrieval effectiveness of the approach to justify their claim.

2.2 Hiemstra's Formulation

Hiemstra (2001) derivation of the query likelihood is based on the direct application of statistical sampling. Sampling is a concept found in most text books on probability theory (Raktoe and Hubert, 1979) and

usually involves examples involving a bag (document) and colored balls (terms). The analogy is as follows: Imagine there is a set of documents $d \in D$, where each document d is represented by a bag of terms. First a document d is selected with probability $P(d)$. Then from that d , a term t_i is selected at random with probability $P(t_i|d)$. The term t_i is recorded and then replaced back into the bag (i.e. sampling with replacement). This step is repeated k times and the output becomes the query $q = \{t_1, \dots, t_k\}$.

Given q , we now ask the IR system which document was most likely to have produced this query. Documents are then ranked according to the joint probability of a query and document, $P(q, d) = P(q|d) \cdot P(d)$. Since the probability of a document $P(d)$ is assumed to be constant, the scoring for each document is approximated by the query-likelihood $P(q|d)$. This is determined by sampling the query terms from each document.

Under this formulation the notion of relevance is not pivotal to the scoring and is in fact ignored all together. Instead of making any claims about relevance in the framework, Hiemstra (2001) seems to suggest that the query likelihood is like a measure of similarity, rather than an estimate of relevance. Consequently, the Language Modeling approach is analogous to the Vector Space Model (Salton and Lesk, 1968), where the similarity between a query and document is used to rank, as opposed to relevance.

2.3 Miller et al's Formulation

Miller et al. (1999) derive the query likelihood approach from a different point of view. By viewing the process of retrieval as a Hidden Markov Model (HMM) they formulate the language model as follows: the observed data, q , is modeled as being the output produced by passing the document through some noisy channel. The analogy is as follows: The noisy channel is the mind of the user, who is believed to have some notion of the ideal document i that they want to retrieve and translates this notion into the query q . Hence, the probability they attempt to estimate is the probability that d was the relevant document i , given that q was produced.

Up until this proposal, relevance was largely ignored or implicitly assumed within the Language Modeling framework. While this approach considers relevance, it only really considers the case when there is only one relevant document and we wish to find that one. While this is seldom the case in practice, Miller et al. (1999) advocate that this is a hypothesis, "Was this the document the user had in mind?". Hence, documents are ranked in decreasing order of

the query likelihood as a means of quantifying how probable this document was like the user's ideal document. Nonetheless, this is a different interpretation of relevance, where Sparck-Jones et al. (2003) argue that this posits that there is one and only one relevant document i . This is an assumption (or limitation) of the model (MLS.A1).

2.4 Lafferty and Zhai's Formulation

An explicit notion of relevance in the Language Modeling framework has been offered by Lafferty and Zhai (2003). They argue that $P(q|d)$ is proportional to the Odds Ratio $O(r|q, d)$ (see Equation 4). We present their arguments below. As with traditional probabilistic modeling the log odds ratio forms the basis of the ranking and is computed by an approximation. In traditional probabilistic models, Bayes' Theorem is applied to $P(r|d, q)$ so that the document probability $P(d|r, q)$ is estimated. For the LM approach Bayes' Theorem is applied to $P(r|d, q)$ such that the focus is on the query probability $P(q|d, r)$, as shown in Equation 5. Mathematically the different decompositions to formulate the traditional and language modeling approaches are equivalent at this point. However, to proceed to the query likelihood from Equation 5, two assumptions are required.

$$\log O(r|d, q) = \log \frac{P(r|d, q)}{P(\bar{r}|d, q)} \quad (4)$$

$$= \log \left(\frac{P(q|d, r) \cdot P(r|d)}{P(q|d, \bar{r}) \cdot P(\bar{r}|d)} \right) \quad (5)$$

$$= \log \frac{P(q|d, r)}{P(q|d, \bar{r})} + \log \frac{P(r|d)}{P(\bar{r}|d)} \quad (6)$$

$$\propto \log \frac{P(q|d, r)}{P(q|\bar{r})} + \log \frac{P(r|d)}{P(\bar{r}|d)} \quad (7)$$

$$\propto \log P(q|d, r) + \log \frac{P(r|d)}{P(\bar{r}|d)} \quad (8)$$

$$\propto \log P(q|d, r) + \log \frac{P(r)}{P(\bar{r})} \quad (9)$$

$$\propto \log P(q|d, r) \quad (10)$$

$$\approx \log P(q|d) \quad (11)$$

LZ.A1 The document and query events d and q are independent given the non-relevance event \bar{r} . Mathematically, this conditional independence assumption is denoted as:

$$P(d, q|\bar{r}) = P(d|\bar{r}) \cdot P(q|\bar{r}) \quad (12)$$

This implies the following:

$$P(q|d, \bar{r}) = P(q|d, \bar{r}) \cdot \frac{P(d|\bar{r})}{P(d|\bar{r})} = \frac{P(d, q|\bar{r})}{P(d|\bar{r})} = P(q|\bar{r})$$

which is applied in Equation 7. Because of this assumption, the document dependent probability $P(q|d, \bar{r})$ reduces to the document independent probability $P(q|\bar{r})$. Since $P(q|\bar{r})$ is constant for all documents, it can be ignored for the purpose of ranking, thus resulting in Equation 8.

LZ.A2 The document and relevance events, d and r , respectively, are assumed to be independent. Mathematically, this independence assumption is captured as follows:

$$P(d, r) = P(d) \cdot P(r) \quad (13)$$

From this assumption we obtain $P(r|d) = P(r)$ by dividing both sides of Equation 13 by $P(d)$, and similarly for the event of non-relevance. Assumption **LZ.A2** is applied in Equation 9. Note, this assumption also implies $P(\bar{r}|d) = P(\bar{r})$.

In Equation 10, the prior probability of relevance over the prior probability of non-relevance is dropped from the ranking because it is constant and independent of the document and query. Finally, Lafferty and Zhai (2003) claim that the Odds Ratio is proportional to the probability of query given the document and the event of relevance. This argument is a much stronger claim than assuming a correlation, one which may not be entirely justifiable. In fact, Robertson (2005) argued that the traditional probabilistic model is not equivalent to the Language model, despite both being derived from the same basis.

3 Analysis of Assumptions

The assumption (**PC.A1**) that the relevance of a document is correlated with the likelihood of the query being generated from that document is a convenient way to simplify the LM approach, which does not appear to be that problematic or radical at first.

Intuitively, we would expect the query terms to be prevalent in the relevant documents, and not so prevalent in non-relevant documents. That is, a good match on query terms *implies*³ relevance (Sparck-Jones et al., 2003). Under the formulation of Ponte and Croft (1998) and Lafferty and Zhai (2003), it is necessary for a correlation between the relevance of a document and the probability $P(q|d)$ to exist (i.e. **PC.A1** and **LZ.A3**). Whilst it has been shown that the correlation between the relevance of a document and the query likelihood is reasonably strong, it has also been shown that it does not always hold in practice (Azzopardi, 2005). Consequently, the PRP can not be guaranteed as the ranking is based on the query

³implies as opposed to infers.

likelihood and not on the relevance of document given the query.

Invoking such assumptions means that relevance within the framework is not explicitly modeled and so it is unclear how relevance information should be handled. Even when relevance was considered by the formulation put forward by Miller et al. (1999) with the guise of the ideal document another problem surfaced stemming from **MLS.A1**; retrieval was focused on finding the one (ideal) relevant document that produced the query. And so finding multiple relevant documents is argued to be inappropriate under this formulation (Sparck-Jones et al., 2003).

The first assumption **LZ.A1** is based on the belief that query terms are only likely from relevant documents and that the query terms are not likely from non-relevant documents. This assumption seems quite reasonable. However, there are instances when this may not be the case. For example, when a term has multiple meanings, then this premise would be violated, and the assumption would not hold.

In Lafferty and Zhai (2003), there is no rationale provided for the second assumption **LZ.A2**. Presumably it was made for mathematical convenience, but it introduces a questionable premise⁴. Dispensing with the dependence between a document and relevance (non-relevance) is inappropriate because this prior will have a significant impact on ranking. Implicitly, the notion of relevance is linked to the document, i.e. either it is relevant or not. In fact, the relevance based language models (Lavrenko and Croft, 2001), estimate the probability $p(d|r)$ in order to rank documents.

Nonetheless, there is one further assumption which needs to be engaged that is more important. The estimation of $P(q|d, r)$ is assumed to be proportional to $P(q|d)$. This approximation means that an implicit assumption (**LZ.A3**) is made, which is which is similar to the assumption **PC.A1**, but instead of a “correlation”, it is an “approximation”.

Since relevance is encoded via the document modeling, there is a greater reliance on obtaining a “good” estimation of the document language model. Ponte (1998) posited that improving the document language model should improve retrieval effectiveness. This decomposition shows why this is a sound intuition, because the generation of terms is reliant on the conditional probability of the query given the document (and it’s assumed relevance). So the only way to af-

⁴We acknowledge that Lafferty and Zhai (2003) are aware that document priors are important to the ranking. However, in the decomposition presented in Lafferty and Zhai (2003), they have dispensed with these priors, in order to obtain the desired outcome.

fect the document’s relevance is via the estimate of the document language model.

The inclusion of relevance within the language modeling approach relevance is acknowledged, but not explicitly modeled. This means that it is difficult to understand how relevance really fits into the model because it is implicitly assumed (and, effectively, ignored, at the operational level).

4 Alternative Formulation

So far, we have presented how relevance has been considered within the different formulations of the LM approach and pointed out some of their shortcomings. The first formulation side stepped the issue of relevance and assumed a correlation existed (**PC.A1**). In the second formulation, there was no assumptions of relevance, but instead suggested the probability of a query given a document as measure of similarity between query and document. The third formulation placed a strict interpretation on the notion of relevance within the model, whilst in the fourth formulation several assumptions are engaged in order to show that the Odds Ratio can be estimated using the probability of a query given a document. But still need to engaged a similar assumption to **PC.A1**, i.e. **LZ.A3**). However, none of these formulation adequately explain the event of (non) relevance within the LM framework. In this section, we present an explanation of relevance within the Language Modeling approach which does not rely on such assumptions or correlations. This formulation is based on the Theorem of Total Probability, and we believe provides a much more satisfactory account of relevance within the LM framework, which is simpler and more intuitive.

4.1 Theorem of Total Probability (TTP)

The theorem of the total probability is as follows: Given an exhaustive space of disjoint events x , where $\sum_x P(x) = 1$ and $\forall x_i, x_j P(x_i, x_j) = 0$, we can express an event probability $P(e)$ by:

$$P(e) = \sum_x P(e|x) \cdot P(x) \quad (14)$$

Apply this theorem to $P(q|d)$, and we obtain:

$$P(q|d) = \sum_x P(q|x, d) \cdot P(x|d) \quad (15)$$

where two disjoint events representing relevance are defined as $x_1 = r$ and $x_2 = \bar{r}$, such that:

$$P(q|d) = P(q|r, d) \cdot P(r|d) + P(q|\bar{r}, d) \cdot P(\bar{r}|d) \quad (16)$$

Equation 16 simply relates the relevance-free query probability $P(q|d)$ to the relevance probabilities. Thus, the probability $P(q|d)$ is composed of two parts. The contribution of the query given the document being relevant $P(q|r, d)$ and the probability of query given the document being non-relevant $P(q|\bar{r}, d)$, weighted by the prior probabilities of relevance and non-relevance given a document, $P(r|d)$ and $P(\bar{r}|d)$, respectively.

Using the theorem of total probability to model relevance within the language modeling framework there is no necessity to resort to making correlations or such strict assumptions about relevance. Intuitively, the decomposition in Equation 16 shows that part of the query generated from the document will be due to its relevance, whilst part of the query generated from the document will be due to its non-relevance. Now, if the prior probability $P(r|d)$ in Equation 16 is set to one, then:

$$P(q|d, r) = P(q|d) \quad (17)$$

This syntactically relates the approximation used in Lafferty and Zhai’s formulation (i.e Equation 10 to Equation 11 in Section 2.4). However, this means that all documents are assumed to be relevant *a priori*. This assumption about the *a priori* relevance of a document, also asserts that the contribution from the non-relevant component is negligible. This means that the assumptions **PC.A1** and **LZ.A3** about the correlation are in fact making a very strong assumption about the relevance of a document.

Another way to derive the transition between Equation 10 to Equation 11 in Section 2.4 is by dividing both sides of Equation 10, by the $P(r|q, d)$. After some manipulation, we obtain:

$$\frac{1}{P(n|d, q)} = P(q|d) \quad (18)$$

This means that positive relevance feedback⁵ is not possible under the model because there is no provision to estimate the relevance, only the non-relevance. This theoretical result confirms the doubt expressed over encoding relevance feedback within the model voiced by Sparck-Jones et al. (2003), where it was argued that relevance feedback did not make sense given the model.

4.2 Applications of TTP

Given the theorem of the total probability and the events relevant and not relevant, there are, in prin-

⁵By relevance feedback, we mean updating the model of relevance, as can be done in the Binary Independence Model (Robertson and Sparck-Jones, 1976), as opposed to query expansion or query term re-weighting.

ciple, two approaches to express the probability $P(q|d) = \prod_{t \in q} P(t|d)$. Approach one is to apply the theorem to term probabilities $P(t|d)$, and approach two is to apply the theorem to the query probability $P(q|d)$.

Micro Approach: Apply total probability to *term* probabilities:

$$\begin{aligned} P(q|d) &= \prod_{t \in q} P(t|d) \\ &= \prod_{t \in q} [P(t|d, r) \cdot P(r|d) + P(t|d, \bar{r}) \cdot P(\bar{r}|d)] \end{aligned} \quad (19)$$

Macro Approach Apply total probability to the *query* probability:

$$\begin{aligned} P(q|d) &= P(q|d, r) \cdot P(r|d) + P(q|d, \bar{r}) \cdot P(\bar{r}|d) \\ &= \left(\prod_{t \in q} P(t|d, r) \right) \cdot P(r|d) \\ &\quad + \left(\prod_{t \in q} P(t|d, \bar{r}) \right) \cdot P(\bar{r}|d) \end{aligned} \quad (20)$$

The first approach views each term probability $P(t|d)$ as a combination of relevance and non-relevance, while the second approach views the query probability $P(q|d)$ as a combination of relevance and non-relevance. In Sections 5.2 and 5.3, we shall show how relevance is explained in the LM framework under these approaches brought about by this alternative formulation.

5 Interpretations and Estimations

In this section, we explain how relevance is interpreted and the probabilities estimated. Before doing so, we first introduce the concept of linear decomposition, which we employ as the basis of our explanation. The application of linear decomposition enables the estimation of the probabilities in the macro and micro approaches.

5.1 Linear Decomposition of $P(t|d, x)$

The approaches in Equations 19 and 20 require an estimate of conditional probabilities of the form $P(t|d, x)$, where x is a place holder for an events $x := r$ and $x := \bar{r}$. Given an event space, where the overlap of d and x is unknown, we apply the following approximation:

$$P(t|d, x) := \delta \cdot P(t|d) + (1 - \delta) \cdot P(t|x)$$

The parameter δ reflects the representative power of a single event d or x . As an extreme, consider t to be distributed in $d \cap x$ as it is distributed in d alone (i.e. $P(t|d, x) = P(t|d)$), then $\delta = 1$.

5.2 Micro Approach: $P(r|d)$ as the mixture parameter

To relate the first approach (decomposition of term probabilities) to LM, we view the probability $P(r|d)$ as the mixture parameter. Under this view, we have, from a mathematical perspective, two options to reach the standard estimate of the query likelihood. Essentially the query likelihood is estimated by the joint probability of the query terms given the document language model $P(q|d)$, where it is typically assumed that query terms are independently and identically drawn from the document. The probability of a term given the document language model, is composed of a two part mixture model. The estimate obtained under the standard query likelihood takes the general form:

$$\begin{aligned} P(q|d) &= \prod_{t \in q} p(t|d) \\ &= \prod_{t \in q} \left(\lambda \cdot \widehat{P(t|d)} + (1 - \lambda) \cdot P(t|c) \right) \end{aligned} \quad (21)$$

where λ is the mixture parameter, $\widehat{P(t|d)}$ is an estimate of the probability of a term in a document (for instance, the maximum likelihood estimate), and $P(t|c)$ is the probability of the term t given the collection c , which is used to not only combat the zero probability problem, but improve the estimate of the document model. Equation 22 defines the general LM approach (Hiemstra, 2001; Miller et al., 1999; Zhai and Lafferty, 2001b) used in practise (which we shall refer to as the standard query likelihood approach).

Now, in order to relate the standard query likelihood estimate to the linear decomposition of relevance within the language model, we have two options. Both yield this estimate of $P(q|d)$, but make different though related assumptions:

Option one: positive Set $P(r|d) := \lambda$ and assume $P(t|d, r) := \widehat{P(t|d)}$ and $P(t|d, \bar{r}) := P(t|c)$. The Figure 1 shows the derivation of how the the standard estimate maps to this option given the Micro Approach.

Option two: negative Set $P(\bar{r}|d) := \lambda$ and assume $P(t|d, \bar{r}) := \widehat{P(t|d)}$ and $P(t|d, r) := P(t|c)$. Note, that the Option two results in the standard estimate of the query likelihood (i.e. Equation 22).

Each option makes certain assumptions about document and the relevance events. The first options expresses the intuition that there is a correlated between

$$\begin{aligned}
P(q|d) &= \prod_{t \in q} \left(\lambda \cdot \widehat{P(t|d)} + (1-\lambda) \cdot P(t|c) \right) \\
&= \prod_{t \in q} \left(P(r|d) \cdot P(t|d) + P(\bar{r}|d) \cdot P(t|\bar{r}) \right) \\
&= \prod_{t \in q} \left(P(r|d) \cdot P(t|d, r) + P(\bar{r}|d) \cdot P(t|d, \bar{r}) \right) \\
&= \prod_{t \in q} \left(P(r, t|d) + P(\bar{r}, t|d) \right) \\
&= \prod_{t \in q} \left(P(t|d) \right)
\end{aligned}$$

Figure 1: Derivation of the standard estimate from the Micro Approach under the Positive option.

a document and relevance (and thus positive), while the second option expresses the opposite. We shall explore each in turn.

Option 1: positive This assumes t to be conditionally independent of r , i.e. it assumes that t is distributed in $d \cap r$ as it is distributed in d . For each term t , we assume $d \cap r = d$, i.e. $P(t|d, r) = P(t|d)$, and $d \cap \bar{r} = \bar{r}$, i.e. $P(t|d, \bar{r}) = P(t|\bar{r})$, where $P(t|\bar{r})$ is approximated by $P(t|c)$. This means that a term is distributed in the conjunction of $d \cap r$ as it is distributed in d alone. If $d \cap r$ is empty (d is not relevant), then the space (evidence) to estimate $P(t|d, r)$ is also empty, which questions the validity of the estimate. If $d \cap r = d$, then we have maximal evidence. From this, we can argue that the assumption is reasonable for a relevant document. The second assumption means that t is distributed in $d \cap \bar{r}$ as it is distributed in \bar{r} overall, where \bar{r} is approximated by the collection c . Essentially, this means that the collection is viewed as a fair approximation of $d \cap \bar{r}$ when looking at statistics of t .

Option 2: negative This assumes t to be conditionally independent of \bar{r} , i.e. it assumes that t is distributed in $d \cap \bar{r}$ as it is distributed in d . For each term t , we assume $d \cap \bar{r} = d$, i.e. $P(t|d, \bar{r}) = P(t|d)$, and $d \cap r = c$, i.e. $P(t|d, r) = P(t|r)$, where $P(t|r)$ is approximated by $P(t|c)$. In option one, d was viewed as a relevant document, but here d is viewed as a non-relevant document, such that $d \cap \bar{r}$ is the evidence space, i.e. the assumption is reasonable if we assume d is not relevant.

Given the two options, which is correct and which assumptions are reasonable? Knowing about the relevance implies the application of $P(t|d, r) = P(t|d)$ for relevant documents, and $P(t|d, \bar{r}) = P(t|d)$ for non-relevant documents. In the first case, $P(\bar{r}|d) \cdot P(t|\bar{r})$ compensates for the lack of knowledge about non-relevance, and in the second case, $P(r|d) \cdot P(t|r)$ compensates for the lack of knowledge about relevance. The collection serves either as an estimate of relevant documents, or as an estimate of non-relevant docu-

ments, depending on the knowledge (feedback) we have or assume for a particular document. Hence, both options are “correct”, where option 1 is for the case of positive relevance feedback and option 2 is for the case of negative relevance feedback.

5.3 Macro Approach: $P(r|d)$ as a meta mixture parameter

Next, we view $P(r|d)$ as a meta mixture parameter, i.e. $P(t|d, r)$ and $P(t|d, \bar{r})$ might be estimated using free mixture parameters (referred to as α and β in the following), that do not need to be related to relevance. Consider the mathematical expression showing the meta-mixture $P(r|d)$, and the mixtures for $P(t|d, r)$ and $P(t|d, \bar{r})$.

$$\begin{aligned}
P(q|d) &= P(r|d) \prod_{t \in q} p(t|d, r) + P(\bar{r}|d) \prod_{t \in q} p(t|d, \bar{r}) \\
P(q|d) &= P(r|d) \prod_{t \in q} (\alpha P(t|d) + (1-\alpha)P(t|r)) \\
&\quad + P(\bar{r}|d) \prod_{t \in q} (\beta P(t|d) + (1-\beta)P(t|\bar{r}))
\end{aligned}$$

Let us consider three options for missing relevance.

Option one: optimistic Set $P(r|d) := 1.0$ where all documents are assumed to be relevant a priori. The derivation shown in Figure 2 shows how standard estimate can be mapped to this option.

Option two: pessimistic Set $P(r|d) := 0.0$, where all documents are assumed to be non-relevant a priori.

Option three: mixed Set $P(r|d) := k$ where $0 < k < 1$, where all documents are assumed to be a mixture of relevant and non-relevant.

Option 1: optimistic Then, in the optimistic case, $P(t|d, r)$ alone affects $P(q|d)$. To estimate $P(t|r)$, we view the collection to correspond to the relevance event, i.e. we set $r := c$ and view c as a token stream of relevant documents (this coincides with the view held by Roelleke and Wang (2006)).

$$\begin{aligned}
P(q|d) &= \prod_{t \in q} (\lambda \cdot \widehat{P(t|d)} + (1-\lambda) \cdot P(t|c)) \\
&= \prod_{t \in q} (\alpha \cdot \widehat{P(t|d)} + (1-\alpha) \cdot P(t|c)) \\
&= 1 \cdot \prod_{t \in q} (\alpha \cdot \widehat{P(t|d)} + (1-\alpha) \cdot P(t|c)) + 0 \cdot \prod_{t \in q} \dots \\
&= p(r|d) \cdot \prod_{t \in q} (\alpha \cdot \widehat{P(t|d)} + (1-\alpha) \cdot P(t|r)) + P(\bar{r}|d) \cdot \prod_{t \in q} \dots \\
&= p(r|d) \cdot \prod_{t \in q} (\frac{P(t|d, r)}{P(q|d, r)}) + P(\bar{r}|d) \cdot \prod_{t \in q} \dots \\
&= p(r|d) \cdot P(q|d, r) + P(\bar{r}|d) \cdot P(q|d, \bar{r})
\end{aligned}$$

Figure 2: Derivation of the Macro Approach under the Optimistic option from the standard estimate.

Option 2: pessimistic In the pessimistic case, $P(t|d, \bar{r})$ alone affects $P(q|d)$. Hence, the probability of a term given non-relevance is approximated by the probability of a term in the collection, i.e. we set $\bar{r} := c$ and view c as a token stream of non-relevant documents. Note, that in both option 1 and 2 lead to an explanation of the standard estimate.

Option 3: mixed In the mixed case, both $P(t|d, r)$ and $P(t|d, \bar{r})$ affect $P(q|d)$. The α is the mixing parameter which adjusts the relevance contribution of the query from the document, and the β is the mixing parameter which adjusts the non-relevance contribution of the query from the document. The setting of these parameters, along with $p(r|d)$ can produce a variety of different estimates for the query likelihood. This provides the opportunity to explore a number of new two stage models.

A special case of option 2 leads to the standard estimate, if the probability of a term given relevance and the probability of a term given non-relevance are approximated by the probability of a term in the collection, i.e. we set $r := c$ and $\bar{r} := c$ and view c as a token stream of relevant and non-relevant documents, simultaneously. Then, if $\alpha = \beta$, $P(q|d, r)$ and $P(q|d, \bar{r})$ will be equal, and since $p(r|d) = 1 - p(r|d)$ the final estimate is equivalent to the standard estimate. However, derived from a completely different view point.

To summarize, we have achieved a framework in which we can interpret the relevance event, where the collection approximates either relevant documents or non-relevant documents, depending on whether we consider positive or negative relevance feedback. This provides an explanations for both cases, i.e. $P(r|d)$ as LM mixture parameter in section 5.2, and $P(r|d)$ as a meta-mixture parameter as considered here, which provide sound explanations of how to view and interpret the relevance event in the LM framework. The explanations are sound for the assumptions we highlighted. The meta-mixture approach is more flexible in the sense that a relevance-independent mixture for

each of the term probabilities $P(t|d, r)$ and $P(t|d, \bar{r})$ is an inherent part of the model.

6 Discussion

In the previous section, we developed approaches to include the relevance event into the language modeling framework. The approaches are based on the Theorem of Total Probability, and we explored different angles of how this theorem could be applied. The consequent usage of the theorem led to two decompositions: One we referred to as the Micro approach, since the decomposition is applied to term probabilities, and the other one, we referred to as the Macro approach, since the decomposition is applied to query probabilities.

We then used linear decomposition to outline how these probabilities could be estimated. We focused on showing how different approaches and options can be approximated to obtain the standard query likelihood estimate. This is because the empirical performance using such an estimates has already be shown to be state of the art (Miller et al., 1999; Hiemstra, 2001; Zhai and Lafferty, 2001b). However, with different probabilities and approximations many different estimates could be obtained, other than this standard estimate. Without this theoretical framework, such extensions would not be possible. An interesting line for future research is to explore the empirical performance of the different estimates which can be derived from these approaches.

Since the focus of this paper has been to account for relevance within the Language Modeling framework, then it is necessary for us to comment on how relevance feedback can be incorporated into the model and how this would fit or change the standard query likelihood estimate. Let's assume that we have evidence such as judged (non)relevant documents, available which we can estimate the probability of a term given (non) relevance. How is this new knowledge encoded within each of the different

interpretations and options? Specifically, given an estimate of $P(t|r)$ or $P(t|\bar{r})$ based on the feedback, how can it be included?

In the Micro Approach, under the Positive Option, the collection is viewed as a stream of relevant documents. Upon receiving positive relevance feedback (i.e. a new estimate of $p(t|r)$), there is no mechanism for updating the query likelihood estimates because of the assumption that $P(t|d, r) = P(t|d)$ (i.e. we have made the assumption that there is no dependency on relevance. As a result, relevance feedback can not be incorporated into the Micro - Positive estimate. However, non-relevance feedback could be included in the model, because $P(t|d, \bar{r}) = P(t|\bar{r})$.

Table 1 summarizes whether $P(t|r)$ or $P(t|\bar{r})$ can be updated under each approach and option taken, when (non) relevance feedback is available. From this table, we can see that only under the Macro Approach, using the Mixed Option, we get a situation where both relevance and non-relevance feedback can be incorporated. Every other option handles either, one form of feedback, or the other. Also, we can see that in the Micro-Positive Approach relevance feedback is not possible, while for the Macro-Positive Approach relevance feedback is possible, despite sharing a similar view about relevant documents. And for the Micro-Negative and Macro-Pessimistic approaches, the former can not incorporate non-relevance feedback, while the later can.

Interestingly, that while we found in Section 3 the theoretical result that positive relevance feedback could not be incorporated within the approach, we see that under different interpretations and options that both forms of feedback can be included. Whether including this information would result in better retrieval performance though is a matter of empirical investigation.

7 Conclusion

The main contribution of this paper is a formalization of relevance within the Language Modeling approach. This provides the basis on which we can understand and interpret relevance within the model without having to resort to make questionable assumptions, such as the independence between the document and relevance (LZ.A2) or relying on assumed correlations (PC.A1, LZ.A3). This frees us from the task of having to argue the semantic meaning (correctness) of the assumptions when wishing to relate LM and the probability of relevance. Instead, the underlying explanation relies upon the application of the Theorem of Total Probability for expressing the

relevance independent query probability $P(q|d)$ as a linear combination $P(q, r|d) + P(q, \bar{r}|d)$. Intuitively, part of the query generated from the document will be due to its relevance, whilst part of the query generated from the document will be due to its non-relevance.

By considering different approaches and options on how to estimate these components we have shown that the standard query likelihood estimate of the Language Modeling approach can be derived in a number of ways. While, this approach has already been extensively tested in the literature (Hiemstra (2001); Ponte and Croft (1998); Miller et al. (1999); Zhai and Lafferty (2001b)), we have also provided the foundations for developing alternative estimates of such Language Models. These new directions include ways of incorporating relevance information in the LM approach in a sound manner. While mathematically sound, we need still to verify its effect on retrieval quality when compared to traditional relevance feedback methods.

In conclusion, this paper covers the theoretical aspects of how the event of relevance can be considered within the LM framework, where our formulation is based on the application of the Theorem of the Total Probability for the relevance event. Future work will focus on an empirical study of this formulation and its interpretations, and on the relationship of relevance in LM and the Binary Independence Retrieval Model.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. David Losada for his helpful comments and feedback.

REFERENCES

- Azzopardi, L. (2005). *Incorporating Context in the Language Modeling Framework for ad hoc Information Retrieval*. PhD thesis, University of Paisley, UK.
- Hiemstra, D. (2001). *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede.
- Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term. In *2002 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 35–41, Tampere, Finland.
- Lafferty, J. and Zhai, C. (2003). Probabilistic relevance models based on document and query generation. In Croft, W. B. and Lafferty, J., editors, *Language Modeling for Information Retrieval*, pages 1–10. Kluwer Academic Publishers.

Table 1: The table lists each approach, option and view that can be taken, and whether it can be mapped to the standard query likelihood and what type of relevance feedback it can incorporate.

Approach	Option	View	Maps to Q.L.	Rel. FB.	Non-Rel. FB.
Micro	Positive	All docs rel.	Yes	No	Yes
	Negative	All docs non-rel.	Yes	Yes	No
Macro	Optimistic	All docs rel.	Yes	Yes	No
	Pessimistic	All docs non-rel.	Yes	No	Yes
	Mixed	Mixed rel/non-rel.	Cond.	Yes	Yes

- Lavrenko, V. and Croft, W. B. (2001). Relevance-based language models. In *Proceedings of the 24th annual international ACM SIGIR conference*, pages 120–127, New Orleans, LA. ACM Press.
- Miller, D. R. H., Leek, T., and Schwartz, R. M. (1999). A hidden markov model information retrieval. In *22nd Annual International ACM SIGIR conference on Research and development in information retrieval*, pages 214–221, California, US. ACM Press.
- Ponte, J. M. (1998). *A Language Modeling Approach to Information Retrieval*. PhD thesis, University of Massachusetts Amherst.
- Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the Twenty First ACM-SIGIR*, pages 275–281, Melbourne, Australia. ACM Press.
- Raktoe, B. L. and Hubert, J. J. (1979). *Basic Applied Statistics*. Marcel Dekker Inc., New York.
- Robertson, S. (2005). On event spaces and probabilistic models in information retrieval. *Information Retrieval*, 8(2):319–329.
- Robertson, S. E. and Sparck-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.
- Robertson, S. E. and Sparck-Jones, K. (1977). The probability ranking principle in ir. *Journal of Documentation*, 33(4):294–304.
- Roelleke, T. and Wang, J. (2006). A parallel derivation of probabilistic information retrieval models. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 107–114. ACM Press.
- Salton, G. and Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36.
- Sparck-Jones, K., Robertson, S. E., Hiemstra, D., and Zaragoza, H. (2003). Language modeling and relevance. In Croft, W. B. and Lafferty, J., editors, *Language Modeling for Information Retrieval*, pages 57–71. Kluwer Academic Publishers.
- Zhai, C. and Lafferty, J. (2001a). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM*, pages 403–410. ACM Press.
- Zhai, C. and Lafferty, J. (2001b). A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 49–56, Tampere, Finland. ACM Press.
- Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22:179–214.