Localized Smoothing for Multinomial Language Models

Victor Lavrenko
Center for Intelligent Information Retrieval
Department of Comupter Science
University of Massachusetts, Amherst, MA 01003
lavrenko@cs.umass.edu

May 2000

Abstract

We explore a formal approach to dealing with the zero frequency problem that arises in applications of probabilistic models to language. In this report we introduce the zero frequency problem in the context of probabilistic language models, describe several popular solutions, and introduce *localized smoothing*, a potentially better alternative. We formulate localized smoothing as a two-step maximization process, outline the estimation details for both steps and present the experiments which show the technique to have potential for improving performance.

1 Overview

Language modeling is quickly gaining recognition as the primary approach to various problems dealing with text. Because language models are estimated from sparse data, many elementary events will have zero probability under the model. In what follows we will briefly introduce the zero-frequency problem in the context of language modeling and outline several popular solutions. We will then propose a possible improvement, based on heuristic techniques that have proven successful in Information Retrieval.

The rest of this report is structured as follows. Section 1.1 describes the task of language modeling and the common unigram formulation for language models. Section 2 reviews the zero frequency problem, which arises when maximal likelihood estimates are used in language models. Section 3 introduces our approach to the zero-frequency problem: *localized smoothing*. Our approach involves a two-step likelihood maximization process, also detailed in Section 3. We evaluate the approach in Section 4, and conclude with important directions for future research in Section 5.

1.1 Language modeling

Language modeling is concerned with estimating how likely it is that a given model M could have generated a sample of text T. In other words, language modeling is an approach to estimating $P(T|M)$. There are a number of approaches to estimating this quantity. Ponte and Croft (Ponte, 1998) assume T to be a binary vector in the vocabulary space with probability of occurrence of every word estimated from M. We take a slightly different approach, assuming T to be a sequence of random variables $T[i] : i = 1...length(T)$, each $T[i]$ takes on the words w in the vocabulary as possible values. We assume that $T[i]$ are independent of each other (a unigram assumption), so the probability of T under M can be rewritten as:

$$P(T|M) = \prod_{i=1}^{length(T)} P(T[i] = w|M)$$

The unigram assumption is also known as *term independence* assumption, and is a common practice in the field of Information Retrieval. It is a known fact that words in the language do not occur independently, for example $P(T[i] = the|T[i-1] = of)$ is much greater than $P(T[i] = the|T[i-1] = the)$, since "of the" is a very common bigram in English, while "the the" is most likely a typo. However, assuming word independence is often necessary to gather sufficient statistics about word occurrence. Furthermore, there is some evidence that preserving word dependencies may not improve the accuracy of probabilistic models of text (e.g. pairwise dependence model by van Rijsbergen (1977)).

Another assumption we make in our model is that $T[i]$ have identical distributions, that is $\forall_{i,j,w} P(T[i] = w) = P(T[j] = w)$. This means we do not model the location of the words in text, we only focus on content. After making this assumption, we can rearrange the terms in the product and group together the tokens $T[i]$ which take on the same value w. The resulting formulation is given below:

$$P(T|M) = \prod_w P(w|M)^{freq(w,T)}$$

2 The zero frequency problem

In this section we look at the estimation of $P(w|M)$ and at the zero frequency problem that may arise in the estimation. In the above formulae $P(w|M)$ means the probability of observing the word w at some position in T, independently of position and all other words occurring in T. The most natural way of estimating this quantity is to use the maximum likelihood of observing w as a sample from M:

$$P_{ml}(w|M) = \frac{freq(w,M)}{length(M)}$$

While this estimate is unbiased, it has a fundamental problem. If M does not contain any instances of the word w, we have $P(w|M) = 0$, which implies $P(T|M) = 0$ for any text sample T that contains w. This becomes a very serious problem when M itself is estimated from a relatively small samples of text (for example from a user's query, as we do in our experimental section). Just because some word w does not occur in the sample from which we estimate M, we cannot assume that w has zero probability under M. This problem is called the zero frequency problem and it is not unique to language modeling. The zero frequency problem arises in numerous applications of probabilistic modeling and bayesian classification, whenever there is insufficient data to form a good model. The problem has been extensively studied in the field of data compression, see (Witten & Bell, 1991) for a prominent example.

2.1 Traditional approaches to the zero frequency problem

There are a number of solutions to the zero-frequency problem, popular in different fields where the problem arises. In the following sections we outline three simple approaches, suggest reasons why they may be deficient and why we may want to seek a better alternative.

2.1.1 Parametric smoothing

One approach to avoiding zero values for $P(w|M)$ is to assume a parametric distribution over the words in the vocabulary, and then fit the parameters of this distribution with the frequency counts from M. There are a number of applicable distributions. Please note that a popular Poisson distribution is not applicable for the task, since if we estimate the mean λ from the model itself, we are still faced with a problem, since the Poisson formulation $\lambda^k e^{-\lambda}/k!$ is still zero whenever $\lambda = 0$. One distribution that is applicable, and used widely is the Gibbs formula:

$$P_{Gibbs}(w|M) = \frac{e^{P_{ml}(w|M)/\tau}}{\sum_v e^{P_{ml}(v|M)/\tau}}$$

This technique is widely known as *softmax* smoothing in Reinforcement Learning and related fields. The formulation avoids zero counts entirely: Gibbs formula allocates a total mass of $1/\sum_v e^{P_{ml}(v|M)/\tau}$ to any word that is not present in M, so it can be viewed as a uniform smoothing technique. The smoothing parameter τ, also referred to as *temperature*, can be used to tune the degree of smoothing.

The method is popular in several fields, but it has two fundamental problems for language modeling. First, it assumes a parametric form of the word distributions. This is a problem, since there have been a few studies indicating that words do not follow simple distributions from the exponential family. Second, Gibbs formulation allocates equal weight to any word that is not found in M. In the next sections we describe two approaches that circumvent these deficiencies.

2.1.2 Uniform smoothing

A simple approach that avoids assumption about the parametric form of word distributions is to simply add a small number ϵ to all probabilities, thus avoiding zero frequencies:

$$P_{uniform}(w|M) = \lambda_M P_{ml}(w|M) + (1 - \lambda_M)\epsilon$$

This is equivalent to assuming a simple mixture model for word generation: with probability λ_M the word is generated by the model M, and with probability $1 - \lambda_M$, the word is generated by a uniform model over the entire vocabulary. If we know the size of our vocabulary $|V|$, we can set $\epsilon = 1/|V|$. However, if the size of the vocabulary is not known, or if we expect new words to enter the vocabulary occasionally, it is a common practice to set $\epsilon = 1/length(M)$, reflecting the fact that we are less and less likely to see new words as our model M gets larger and larger (this is related to Zipf's law of word occurrences).

This method is very simple, and has the advantage of not assuming parametric distributions for word occurrences. However, it suffers from assigning equal probabilities to all words that do not occur in M. This means that if neither of the words "the" and "Zipf" are in M, the uniform smoothing method will assign equal likelihoods to their occurrence in T. This is a problem, since we know that “the” is a very common word, and so is likely to occur in any piece of text, while “Zipf” is certainly not. The following method resolves this deficiency.

2.1.3 Smoothing with a prior (global smoothing)

We can avoid the problem highlighted in the previous section by selecting a better mixture model in place of the uniform model. A natural choice is to substitute the prior probability $P(w)$ instead of ϵ for every word:

$$P(w|M) = \lambda_M P_{ml}(w|M) + (1 - \lambda_M)P(w)$$

This will have the desired effect of closer matching the specifics of language, giving higher likelihood to observing “the” than to observing “Zipf”. The prior probability $P(w)$ is estimated from the universe of all English texts, in practice this means as large a collection of texts as we can get a hold of. It is worth mentioning that a common practice is to smooth the prior probability $P(w)$ with a uniform model, to avoid the possibility of some words missing from our large collection (however the weight allocated to the uniform model is much smaller than the weight allocated to $P(w)$).

This formulation has been rather successful in applications of language modeling, but there is still room for improvement. We may observe that a global model may not be the best fit for the mixture. This is due to the fact that a global model gives the true prior probability $P(w)$ for occurrence of w in a *random* piece of text. If we are considering documents in a narrow domain, this distribution may be a poor fit. For instance, the word “Zipf” is a fairly common word in the Information Retrieval literature, whereas in a random piece of text the prior probability of observing “Zipf” is virtually nonexistent. This discrepancy leads us to examine approaches that model the context of our model M, which we do in the next section.

3 Localized Smoothing

In this section we introduce the main contribution of this work: an approach to localized smoothing: mixing the model M with its context. We will refer to the context as the *zone* of M. We assume a similar mixture model: with probability λ_M the word w is generated by the original maximum likelihood model of M, with probability λ_{zone} it is generated by the contextual model of M, and with probability λ_{global} the word is generated by the global model of word occurrences:

$$P(w|M) =$$

$$= \lambda_M P_{ml}(w|M) + \lambda_{zone} P(w|M_{zone}) + \lambda_{global} P(w)$$

The motivation is as follows. When estimating the likelihood of T, we assume that salient words will come from M, related concepts and synonyms will be generated by contextual model, and the functional words will be generated by the prior (global) model. Now we turn our attention to estimating the zone model of M.

First, we have to identify the zone of a model M. The zone is the projection of M onto the space of text samples, some of which are expected to contain the context of M. In probabilistic terms, we define the zone of M to be the subset of text samples that maximizes the posterior likelihood of M being their source:

$$zone(M) = \arg \max_{k,\{T_1...T_k\}} P(M|\{T_1...T_k\})$$

Note that the size k of this subset $\{T_1...T_k\}$ is not specified, and is a variable in maximization. Once we have determined the zone of M, we can estimate the model of that zone. We define the zone model to be the model that maximizes the probability of observing the set of samples $\{T_1...T_k\}$:

$$M_{zone} = \arg \max_{M'} P(\{T_1...T_k\}|M')$$

Note that the two maximization steps are distinct: in the first step we are searching over all subsets of text samples, while in the second we are searching in the space of models. We now turn to the details of estimation in each step.

3.1 Maximum likelihood context zone of M

We defined the zone to be a subset $\{T_1...T_k\}$ of our space of text samples, which maximizes the posterior probability of the model M being its source. We can use Bayes theorem to express this posterior as the ratio of probability of $\{T_1...T_k\}$ under the model M over the prior probability for $\{T_1...T_k\}$. Note that the prior probability for M drops out because it is a constant in the maximization step:

$$zone(M) = \arg \max_{k,\{T_1...T_k\}} P(M|\{T_1...T_k\})$$

$$= \arg \max_{k,\{T_1...T_k\}} \frac{P(\{T_1...T_k\}|M)}{P(\{T_1...T_k\})}$$

Note that we must use a smoothed version of $P(T_i|M)$ in the numerator, as warranted in Section 2. We assume smoothing with a prior model, but any other smoothing could be used. Because $\{T_1...T_k\}$ is the set of independent text samples, we can rewrite the joint probability as the product of the marginals as follows:

$$zone(M) = \arg \max_{k,\{T_1 \ldots T_k\}} \prod_{i=1}^{k} \frac{P(T_i|M)}{P(T_i)}$$

The formulation above suggests an simple composition of the zone. Observe that the product is maximized as long as the individual terms in the product each exceed 1. Therefore the zone may be composed of text samples T_i which are more likely under the model M than they are likely a-priori, in our universe of text samples. We may, for reasons suggested in Section 4, wish to further constrain the zone to the samples T_i which have the likelihood ratio exceeding $\theta > 1$:

$$zone(M) = \{T_i : \frac{P(T_i|M)}{P(T_i)} > \theta \geq 1\}$$

Now that we have defined the composition of the zone of M, we turn our attention to estimating the maximum likelihood model for that zone.

3.2 Maximum likelihood model of the context zone

We defined the zone model M_{zone} to be the model that maximizes the likelihood of observing $\{T_1 ... T_k\}$:

$$M_{zone} = \arg \max_{M'} P(\{T_1 ... T_k\}|M')$$

Note that M_{zone} as defined above also maximizes its own posterior likelihood $P(M'|\{T_1 ... T_k\})$ when $P(M')$ is uniform over M'. As before, because $\{T_1 ... T_k\}$ is the set of independent samples, we can decompose the joint probability into the product of the marginals:

$$M_{zone} = \arg \max_{M'} \prod_{i=1}^{k} P(T_i|M')$$

We employ our definition of $P(T|M)$, as specified in Section 1.1 to obtain:

$$M_{zone} = \arg \max_{M'} \prod_{i=1}^{k} \prod_{w} P(w|M')^{freq(w,T_i)}$$

Now we observe that we can re-arrange the ordering of terms in the product:

$$M_{zone} = \arg \max_{M'} \prod_{w} \prod_{i=1}^{k} P(w|M')^{freq(w,T_i)}$$

Observe that $P(w|M')$ is independent of i, so we can transform the product of exponents to an exponent of the sum:

$$M_{zone} = \arg \max_{M'} \prod_{w} P(w|M')^{\sum_{i=1}^{k} freq(w,T_i)}$$

Logarithm is a non-decreasing transformation, so we can take the logarithm of the above expression without any affect on maximization. We also bring the sum to the outside of the logarithm, and use negation to change maximization to minimization:

$$M_{zone} =$$

$$= \arg \min_{M'} - \sum_{w} \left(\sum_{i=1}^{k} freq(w,T_i) \right) \log P(w|M')$$

The next step is to note that minimization is unaffected if we multiply the objective by a constant. Since $\sum_{i=1}^{k} length(T_i)$ is a constant ($\{T_1 ... T_k\}$ is fixed and we are maximizing over M'), we can write:

$$M_{zone} =$$

$$= \arg \min_{M'} - \sum_{w} \left(\frac{\sum_{i=1}^{k} freq(w,T_i)}{\sum_{i=1}^{k} length(T_i)} \right) \log P(w|M')$$

However, the term before the logarithm is simply the maximum likelihood estimate for w under the set $\{T_1 ... T_k\}$:

$$M_{zone} = \arg \min_{M'} - \sum_{w} P_{ml}(w|\{T_1 ... T_k\}) \log P(w|M')$$

Observe that the objective function in the minimization above is exactly the relative entropy between $\{T_1 ... T_k\}$ and M'. Entropy is minimized when two distributions are identical, which leads us to a simple and intuitive formulation for M':

$$P(w|M') = P_{ml}(w|\{T_1 ... T_k\})$$

3.3 Mixing probabilities

We have detailed both maximization steps involved in forming a context model. One practical issue remains: selection of appropriate mixing weights λ_M, λ_{zone} and λ_{global}. An obvious constraint is that weights sum to one. There exist a number of approaches to selecting the weights so as to maximize some objective, such as log likelihood of the training data (e.g. the use of EM algorithm by T. Hoffman in his work on global topic mixtures). At this point in our research we opt for using a closed-formed estimates derived by (Witten & Bell, 1991):

$$\lambda_M = \frac{\sum_{w \in M} freq(w,M)}{\sum_{w \in M}(1 + freq(w,M))}$$

$$\lambda_{zone} = \frac{\sum_{w \in zone} freq(w, zone)}{\sum_{w \in zone}(1 + freq(w, zone))}$$

To ensure the weights sum to one, we employ nesting as follows:

$$P(w|M) = \lambda_M P_{ml}(w|M) + (1 - \lambda_M) *$$

$$* \left[\lambda_{zone} P(w|M_{zone}) + (1 - \lambda_{zone}) P(w)\right]$$

Note that in our experiments $P(w)$ is further smoothed by a uniform model, as described in Section 2.1.3.

4 Experiments

In this section we describe an implementation of the localized smoothing approach described in Section 3. We test our formulation against a popular approach of smoothing with a prior (global smoothing), detailed in Section 2.1.3. Aside from smoothing, we keep all modeling details exactly the same for both approaches. We compare the effectiveness of the two approaches on the TREC ad-hoc retrieval task (Allan, Callan, Feng, & Malin, 1999).

We use Detection Error Tradeoff (DET) curves (see Figure 1) to evaluate the impact of smoothing approaches on retrieval performance. DET curves are used extensively in signal detection literature and have several advantages over the traditional Recall-Precision curves used in the Information Retrieval community. The motivating factor for choosing DET curves over Recall-Precision curves in this evaluation is that DET curves are less influenced by “richness” (the a-priori probability of on-target item in the dataset). For a more detailed description of DET curves, see (Martin, Doddington, Kamm, & Ordowski, 1997).

4.1 Experimental setup

The ad-hoc retrieval task is the task of ranking a collection of documents by their estimated relevance to the user’s query. We use a set of 50 queries (title versions of TREC queries 251-300). Each query consists of 3-4 words on average (typical for web queries). We use the AP’1988 collection of newswire articles as our data set. The collection contains around 80,000 documents, the average document length is around 300 words. Out of the original 50 queries, 48 queries were judged by TREC assessors to have relevant documents in the dataset. The number of relevant documents ranged from 1 to 280, with an average of 35.

Our experiments take the following form. For each of the 48 queries, we form a maximum-likelihood model M^q_{ml} from the words in the query. The models M^q_{ml} are extremely sparse, since queries contain very few words. We smooth each model M^q_{ml} either globally (Section 2.1.3), or using our *localized* approach (Section 3), to obtain a smoothed version M^q. Then, for each document D in the collection, we compute the posterior likelihood that the smoothed model M^q is the source from which D was generated: $P(M^q|D)$. We consider this posterior to be an estimator of the degree of relevance to the query and rank the documents by $P(M^q|D)$.[1]

4.2 Effect of zoned smoothing

In our first experiment we set the zoning threshold θ to 1 (Section 3.1), corresponding to true maximization. Figure 2 shows the distributions of document scores (log-likelihood ratios) for documents that were judged relevant and non-relevant by TREC assessors. The distributions are pooled across all 48 queries, and so are biased towards queries with more relevant documents. The distributions were constructed using a non-parametric kernel density estimator[2].The left graph shows the distributions obtained using global smoothing, the right half presents the results of our localized approach.

One thing that becomes immediately apparent from looking at the distributions is the increase in the variance of scores when using localized approach. However, the variances did not increase proportionally: with global smoothing the variance of non-relevant documents was lower than that of the relevant ones, but with localized smoothing the variance of non-relevant documents became much higher than that of the relevant documents. Another peculiarity of the localized smoothing approach is the small bump in the density of non-relevant documents at the high scores. This suggests that localized approach produced a number of very highly-ranked non-relevant documents. This would be a discouraging observation if our goal was to produce a low-recall high-precision system.

Figure 3 displays the same information as Figure 2, only in a form of a DET curve, allowing easier analysis. We see that localized feedback indeed shows inferior performance at low recall (high miss rate). We also clearly see the effect of increased variance in the distribution of non-relevant scores. On a DET curve, increased variances of non-relevant scores translate to steeper curves (see

[1] Actually, we rank by log $P(D|M^q)/P(D)$, which is seen to be equivalent after transforming the posterior using Bayes theorem and noticing that the prior $P(M^q)$ does not affect the ranking of documents

[2] We used gaussian kernels with automatic bandwidth selection

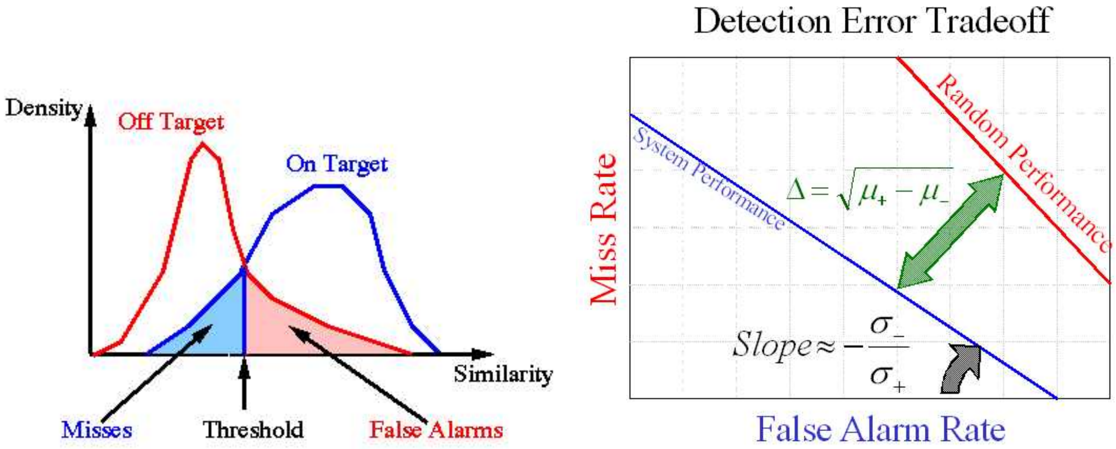

Figure 1: DET curves are a way to visualize tradeoff between misses and false alarms. Left: distributions of on-target and off-target scores, shaded areas under the curves correspond to miss and false alarm errors. Right: corresponding DET curve, obtained by varying the threshold from $-\infty$ to ∞. *NOTE: on a DET curve lower means better.*

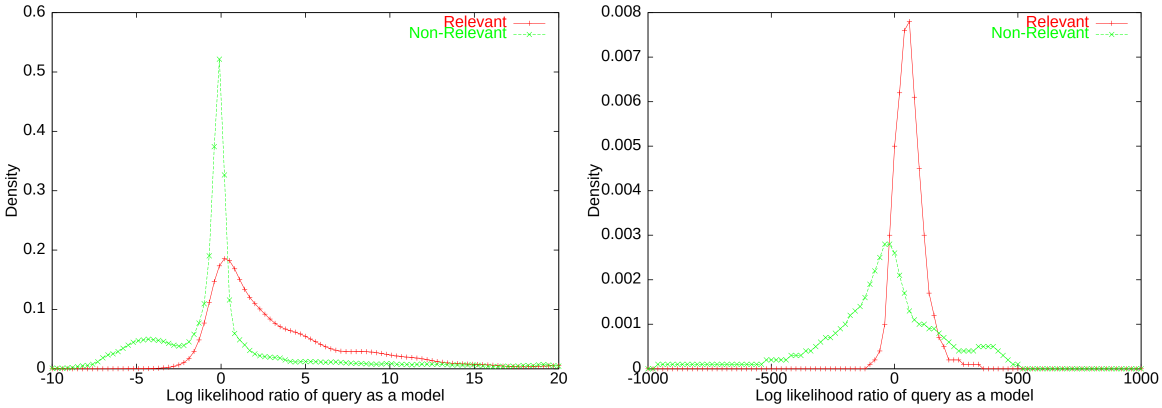

Figure 2: Average distributions of scores of relevant and non-relevant documents. Left: global smoothing. Right: localized smoothing.

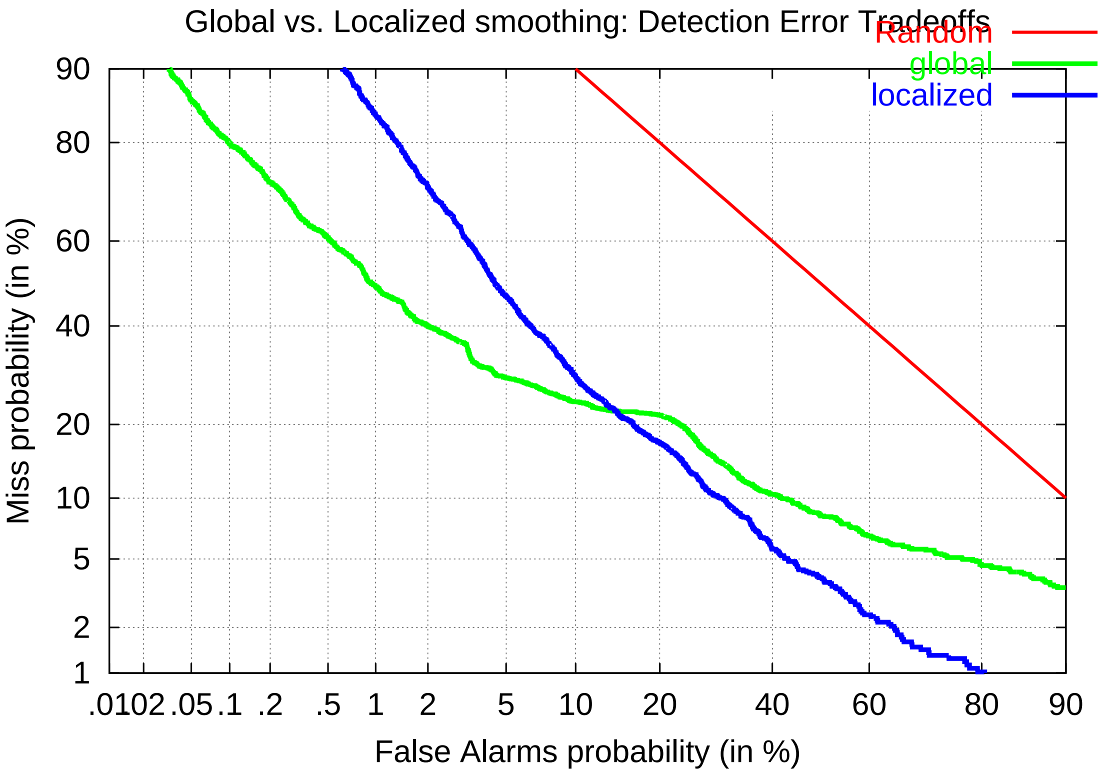

Figure 3: Performance of global vs. localized smoothing at $\theta = 1$.

Figure 1).

The increased variance of non-relevant documents makes sense. When we use a global smoothing approach, all documents that do not contain query words are assigned their prior in the collection as a whole, which is uniformly very low. When we use a zoned model for smoothing, a large class of documents that are nearby to the zone receive a significant boost in their scores, whereas documents that are not nearby are assigned a fraction of their global prior, driving their scores even lower. From Figure 3, we see that increased variance appears to have a detrimental effect overall – resulting in better performance only at high levels of recall (miss rate below 20%). Improvements in that region are generally ignored by researchers in IR, though other fields (e.g. TDT) may consider the improvement useful.

4.3 Impact of thresholding the zone

The detrimental effect of very high variance leads us to consider tightening the context zone around M^q_{ml} (Section 3.1). To do this, we perform a number of experiments setting θ to various values and observing the effect it has on performance. We experimented with θ taking values $1, e, e^2, e^4$. Note that successively larger values of θ translate to smaller and smaller sizes of the context zone around M^q_{ml}.

The results are presented in Figure 4. We observe that increased values of θ indeed result in improved performance in the low-recall region. For example, setting $\theta = e^4$ results in a consistent five-fold improvement in False Alarm rate at low levels of recall. However, the performance rapidly gets worse at higher recall. Setting threshold to $\theta = e$ appears to give reasonable results overall, and we compare this setting to the performance of global smoothing in Figure 5. We observe that performance in the low-recall range is still worse, but not nearly as much as with $\theta = 1$. However, for higher recall, localized smoothing gives dramatic improvements reducing the false alarm rate 5 to 6 times. This is an interesting improvement from the standpoint of evaluating the technology, even if it does not translate to higher relevance of top-ranked documents.

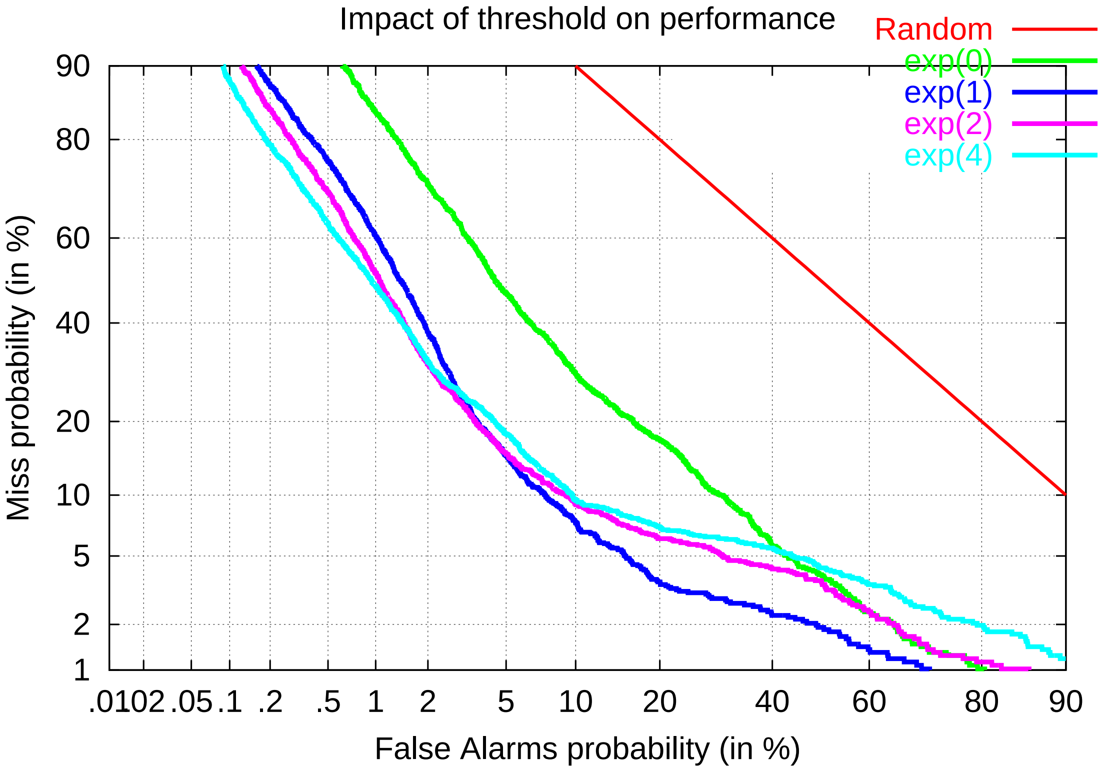

Figure 4: Effects of threshold θ on performance of localized smoothing.

5 Directions for future work

We believe the technique of localized smoothing presented in this paper has significant potential for improving the quality of language models. Our evaluation demonstrated that simple maximization of the posterior does not produce an optimal context zone for the original model (our approach performed poorly for $\theta = 1$). We intend to investigate alternative optimization procedures for finding the optimal query zone. We also need to investigate how this approach compares to smoothing used in other language modeling formulations (e.g. Ponte, 1998).

6 Conclusions

We presented a novel approach to localized smoothing of language models, based on modeling the context zone around the original model. Our technique relies on a two-step likelihood maximization, which is detailed in Section 3. We tested our approach against a commonly used global smoothing technique on a standard set of TREC queries. Experiments show that our approach provides significant improvements in the high-recall region, but results in decreased quality at the top of the ranked list.

7 Acknowledgments

This material is based on work supported in part by the Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, and in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

Allan, J., J. Callan, F. Feng, & D. Malin (1999). INQUERY and TREC-8. In D. Harman (Ed.), *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*.

Martin, A., G. Doddington, T. Kamm, & M. Ordowski (1997). The det curve in assessment of detection task performance. In *EuroSpeech*, pp. 1895–1898.

Ponte, J. (1998). *A Language Modeling Approach to Information Retrieval*. Ph. D. thesis, Dept. of Computer Science, University of Massachusetts, Amherst.

van Rijsbergen, C. J. (1977). A theoretical basis for the

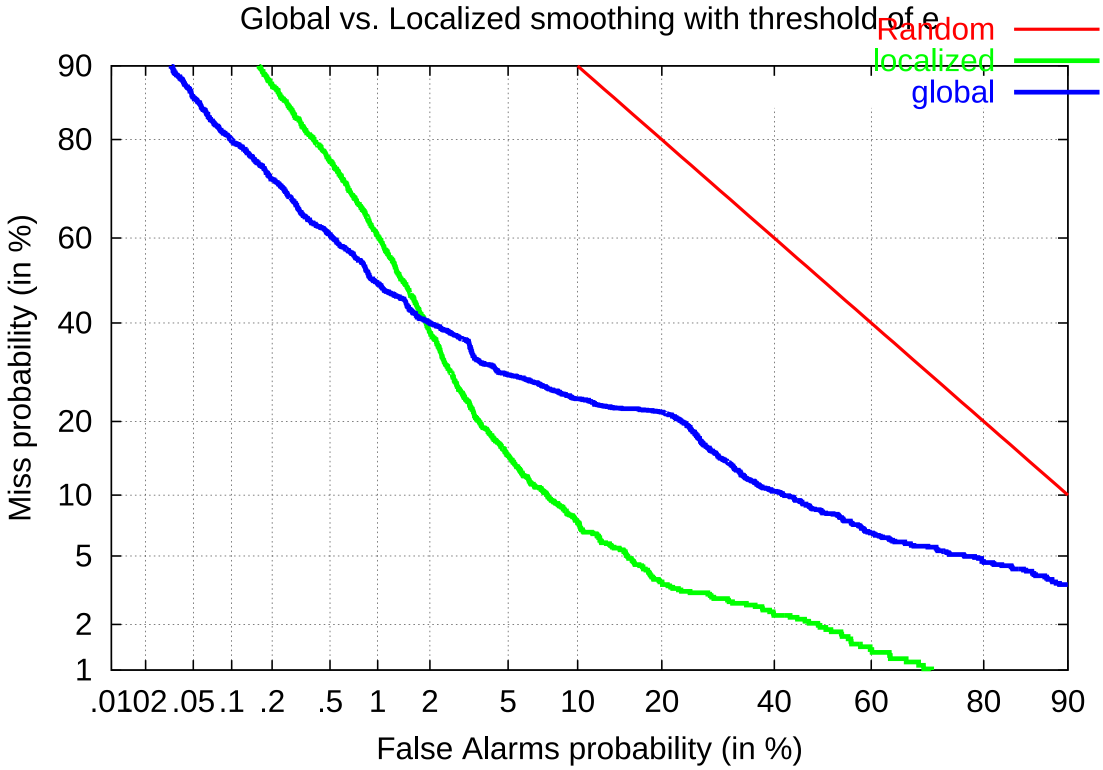

Figure 5: Performance of global vs. localized smoothing at $\theta = e$.

use of co-occurrence data in information retrieval. *Journal of Documentation 33*, 106–119.

Witten, I. H. & T. C. Bell (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Information Theory 37*, 1085–1094.