Chapter 1

LANGUAGE MODELS FOR TOPIC TRACKING (AUTHOR VERSION)

The importance of score normalization

Wessel Kraaij
TNO TPD
P.O. Box 155, 2600 AD Delft, The Netherlands
kraaij@tpd.tno.nl

Martijn Spitters
TNO TPD
P.O. Box 155, 2600 AD Delft, The Netherlands
spitters@tpd.tno.nl

Abstract Generative unigram language models have proven to be a simple though effective model for information retrieval tasks. In contrast to ad-hoc retrieval, topic tracking requires that matching scores are comparable across topics. Several ranking functions based on generative language models: straight likelihood, likelihood ratio, normalized likelihood ratio, and the related Kullback-Leibler divergence are evaluated in two orientations. Best performance is achieved by the models based on a normalized log-likelihood ratio. Key component of these models is the a-priori probability of a story with respect to a common reference distribution.

Keywords: Language Models, Information Retrieval, Score Normalization

1. Introduction

Topic tracking is one of the tasks of the annual Topic Detection and Tracking (TDT) evaluation workshop, which was first organized in 1996.

Main purpose of the TDT project is to advance the state-of-the-art in determining the topical structure of multilingual news streams from various sources, including newswires, radio and television broadcasts, and Internet sites. See (Wayne, 2000) for a detailed overview of the TDT project. The tracking task models the information-need of a user who hears about a certain event on the radio or television and wants to be notified by all follow-up stories in a number of pre-specified information sources in different languages. TDT is challenging because it combines several problems: automatic speech recognition and segmentation of continuous media like radio and television, cross-lingual access to data and a topic tracking task without supervised relevance feedback. A topic tracking system is initialized with one or a few stories describing a certain news event, and must track this topic in a stream of new incoming stories. The most basic form of a tracker makes just binary decisions: a story is on-topic or off-topic. In practice such a decision is based on thresholding a score which is designed to be some monotonic function of the probability that the story is on-topic.

The goal of this study is to investigate whether generative probabilistic models that have been successfully applied to ad-hoc IR tasks (Hiemstra, 1998; Hiemstra and Kraaij, 1999; Kraaij et al., 2000) can be applied to the tracking task as well. The tracking task contains an additional difficulty in comparison with the ad-hoc task because it is required that scores are comparable across topics since the decision threshold is equal across topics. In this paper, we will review several ways to use generative models for tracking and methods to obtain comparable scores across topics. We hope to find a single model which is effective for both the ad-hoc and tracking task.

The remainder of this paper is organized into three main sections. Section 2 discusses different lay-outs for the use of language models for ad-hoc IR and topic tracking. In particular, we will look at model-internal and external normalization methods. In Section 3 we describe experiments with a selection of models that we carried out on the TDT development data and on TREC-8 Ad-hoc data. We conclude the paper with a discussion and conclusions.

2. Language models for IR tasks

The basic problem underlying most information retrieval problems is that of ranking documents based on relevance with respect to a certain information need. This need could be an ad-hoc query, a long-standing topic of interest or - in a more dynamic fashion - an event of interest. For this class of problems, models have to impose an ordering on documents based on their supposed relevance: the probability ranking principle (Robertson, 1977). An implicit constraint is that these models need to be able to cope with documents of different lengths. In some of the TREC collections for example, document sizes can differ with several orders of magnitude. If a score would be correlated with document length this would cause highly inflated scores for long documents. For another class of IR problems, which is more related to classification, simple ordering is not enough. Here we need to be able to interpret scores in an absolute way, since the score is used to classify a document as relevant or not relevant. This is for example the case in filtering applications. In the TREC adaptive filtering task, the decision threshold can be adjusted for each topic on the basis of relevance feedback. However, in the TDT tracking task the decision threshold is taken to be uniform across all topics. This makes sense, since the task does not allow any supervised relevance feedback. As a consequence scores must be comparable across stories (documents) and topics (queries). For certain applications (e.g. document clustering) it is even desirable that matching scores fulfill another constraint, namely symmetry (Spitters and Kraaij, 2002).

2.1 Score properties of probabilistic models

It is instructive to review the relationship of probabilistic models for IR with regard to the aspect of score normalization. For reasons of legibility, we will present the models from the point of view of an ad-hoc IR problem, i.e. we talk about documents and queries. In most cases (unless stated otherwise) these models also apply to the tracking task, after replacing the query by a topic and documents by stories.

Following (Sparck Jones et al., 2000) we define:

- Q is the event that the user has a certain information need and describes it with description Q (In a tracking setting: The user is interested to track stories related to the topic described by T)
- D is the event that we consider a document with description D (Tracking setting: we are considering a story S)
- L is the event that D is liked (or relevant). (L is the event that S is liked)
- $\bar{L}$ is the event that D is not liked (or relevant). ($\bar{L}$ is the event that S is not liked). to T)

Now, for a certain query, we want to rank documents on the probability that they are liked. This can be done by estimating $P(L|D,Q)$: the probability that a document is liked given its description and the description of the query. In order to simplify further computation[1], documents are ordered on log-odds of being liked, which is an order preserving operation.

2.1.1 Document likelihood. The classical next step is to apply Bayes' rule in order to express the matching score based on log-odds in terms of $P(D|L,Q)$ i.e. the probability that a document is described by D_i when we know it is relevant for a certain query Q. This model describes the situation where we have one query and several documents.

$$\log \frac{P(L|D_i,Q)}{P(\bar{L}|D_i,Q)} = \log \frac{P(D_i|L,Q)}{P(D_i|\bar{L},Q)} + \log \frac{P(L|Q)}{P(\bar{L}|Q)} \tag{1.1}$$

Since we do not have any information about the prior probability of relevance given a certain query, we assume a uniform prior so the second term in (1.1) can be dropped for ranking purposes. Ranking is then solely based on the log-likelihood ratio $P(D_i|L,Q)/P(D_i|\bar{L},Q)$. One could interpret this log-likelihood ratio as follows: "How likely is the description of document D_i if we assume the document is relevant to Q?". This likelihood is normalized w.r.t. a model based on descriptions of non relevant documents[2]. Because the model is about one query and several documents, scores are inherently comparable across documents, due to the normalizing denominator likelihood.

Now $P(D_i|L,Q)$ (and $P(D_i|\bar{L},Q)$) can be estimated in various ways. In the Binary Independence Model (Robertson and Sparck Jones, 1976) also known as the Binary Independence Retrieval (BIR) model(Fuhr, 1992), D is described as a vector of binary features x_k, one for each word in the vocabulary. Further development of the log-odds assuming term independence leads to the classical Robertson-Sparck-Jones formula for term weighting. Estimation of $P(D_i|L,Q)$ is usually based on the assumption that there is prior knowledge about some relevant documents. Such a situation is essentially equivalent to supervised text classification based on the Naive Bayes assumption, where the classes are $L, \bar{L}$ (Lewis, 1998). In the absence of relevance information, the BIR model reduces to *idf* term weighting, which is quite weak. The matching score based on the log-likelihood is basically a sum of term weights over all terms in the vocabulary, but usually it is assumed that $P(x_k|L,Q) = P(x_k|\bar{L},Q)$ for all terms that do not occur in the query. This means that scores of the 'typical' BIR model are comparable for documents but not comparable for queries, since scores depend on the query length and not on document length. The BIR model can thus be used unchanged for the ad-hoc IR task, but scores have to be normalized for topic length, if we would want to use the BIR model for tracking.

One can also estimate $\frac{P(D_i|L,Q)}{P(D_i|\bar{L},Q)}$ in a generative framework where D_i is defined as a sequence of terms. In such a generative framework we can think of $P(D_i|L,Q)$ as the probability that D_i is generated as a sequence of terms from a unigram language model M_R which is constrained by Q and L i.e. which describes the probability of observing words in documents relevant to Q. As usual, term independence is assumed. This particular model is also referred to as 'document-likelihood' (Croft et al., 2001a). In a similar way we can think of $P(D_i|\bar{L},Q)$ as the probability that D_i is generated from a model estimated on non relevant documents, which we can approximate by a model of the collection: $P(w|M_{\bar{R}}) \approx P(w|M_C)$. Since the vast majority of documents are not relevant, this seems a reasonable assumption. Substituting the generative estimates in the log-likelihood ratio results in:

$$\log \frac{P(D_i|L,Q)}{P(D_i|\bar{L},Q)} \approx \sum_{w \in D_i} d_w \log \frac{P(w|M_R)}{P(w|M_C)} \tag{1.2}$$

where d_w is the term frequency of the word w in the document. Just like the BIR Model, it is difficult to estimate $P(w|M_R)$ for ad-hoc queries in the absence of relevance information. Applying maximum likelihood estimation on a short query would yield a very sparse language model. However, recently a new estimation technique has been developed to estimate $P(w|M_R)$ in a formal and effective way(Lavrenko and Croft, 2001). The so-called Relevance Model is based on estimating the joint distribution P(w,Q) by making use of term cooccurrence in the document collection. For tracking, estimation is easier, since there is at least one example story. Stories are usually considerably longer than the typical ad-hoc query.

Regarding score comparability, the situation is reversed with respect to the BIR model. Scores are independent of query length (a relevance model is a probability distribution function over the complete vocabulary) but are dependent on the length of the generated text, as can be seen in formula (1.2). We can illustrate this by comparing the scores of a document A and a document B, which consists of two copies of document A. Intuitively, both documents are equally relevant, but this is not reflected in the score. A simple correction is to normalize by document (story) length, making the score usable for ad-hoc and tracking tasks. Interestingly, a ratio of length normalized generative probabilities can also be interpreted as a difference between cross entropies:

$$\sum_w \frac{d_w}{\sum_w d_w} \log \frac{P(w|M_R)}{P(w|M_C)} = \sum_w P(w|M_{D_i}) \log P(w|M_R) - \sum_w P(w|M_{D_i}) \log P(w|M_C) \tag{1.3}$$

Here M_{D_i} is a unigram model of document D_i, which is constructed on the basis of maximum likelihood estimation. The basic ranking component in (1.3) is the (negated) cross-entropy $H(M_{D_i}; M_R)$, which is normalized by the cross-entropy $H(M_{D_i}; M_C)$. We will refer to this length normalized likelihood ratio with the shorthand $NLLR(D; Q, C)$.

2.1.2 Query Likelihood. Coming back to the original log-odds model, Bayes' rule can also be applied to derive a model where the log-odds of being relevant is described in terms of $P(Q_j|L, D)$, i.e. the probability that a query is described by Q_j when we know that a document described by D is relevant (Fuhr, 1992; Lafferty and Zhai, 2001).

$$\log \frac{P(L|D, Q_j)}{P(\bar{L}|D, Q_j)} = \log \frac{P(Q_j|L, D)}{P(Q_j|\bar{L}, D)} + \log \frac{P(L|D)}{P(\bar{L}|D)} \tag{1.4}$$

Strictly spoken, this model describes the situation where there is one document and a number of queries submitted to the system.Still, the model can be used for document ranking provided that the document models are constructed in a similar manner and do not depend on document length. This time, the likelihood ratio computes how typical the query description Q_j is for document D in comparison to other query descriptions. Key element for the comparability of scores of different queries is the normalizing denominator $P(Q_j|\bar{L}, D)$. Again, there are multiple ways to estimate $P(Q_j|L, D)$. A query representation by a binary feature vector leads to the Binary Independence Indexing (BII) model (Fuhr, 1992), which is closely related to the first formulated probabilistic IR model of Maron and Kuhns (Maron and Kuhns, 1960). Because of estimation problems, these models have to our knowledge not been used for practical IR-tasks like ad-hoc queries or tracking. With regard to score comparability, these models should be normalized for query length in order to be used for tracking, the models can be used unchanged for ad-hoc tasks.

The query-likelihoods in (1.4) can also be estimated in a generative framework. We could think of $P(Q_j|L, D)$ as the probability that Q is generated as a sequence of terms from a model which is constrained by D and L. This means, we have a document with description D, which we know is relevant, so we can use this document as the basis for a generative language model and calculate the query likelihood. Analogously to the document-likelihood model (1.2), we assume term independence and approximate $P(Q_j|\bar{L}, D)$ by the marginal $P(Q)$:

$$\log \frac{P(L|D, Q_j)}{P(\bar{L}|D, Q_j)} = \sum_{w \in Q_j} q_w \log \frac{P(w|M_D)}{P(w|M_C)} + \log \frac{P(L|D)}{P(\bar{L}|D)} \tag{1.5}$$

where q_w is the number of times word w appears in query Q. The prior probability of relevance should not be dropped this time, since it can be used to incorporate valuable prior knowledge, e.g. that a webpage with many inlinks has a high prior probability of relevance (Kraaij et al., 2002). This model is directly usable for the ad-hoc task, since scores are comparable across documents of different lengths, due to the maximum likelihood procedure, which is used to estimate $P(w|M_D)$. Usually, the denominator term $P(Q_j|\bar{L}, D)$ is dropped from the ranking formula, since it does not depend on a document property.

$$\log P(L|D, Q_j) = \log P(L|D) + \sum_{w \in Q_j} q_w \log \frac{P(w|M_D)}{P(w|M_C)} \tag{1.6}$$

This results in the basic language modeling approach (1.6) as formulated in (Hiemstra, 1998) and (Miller et al., 1999).

If we want to use model (1.5) for tracking, scores should be comparable across queries, therefore the denominator, which depends on the query, should not be dropped. In addition, scores have to be normalized for topic (=query) length^3, which leads again to a ranking formula consisting of the difference between two cross entropies (for simplicity, we assume a uniform prior and drop the prior odds of relevance term in (1.5)):

$$\sum_w \frac{q_w}{\sum_w q_w} \log \frac{P(w|D)}{P(w|M_C)} = \\ \sum_w P(w|M_{Q_j}) \log P(w|M_D) - \sum_w P(w|M_{Q_j}) \log P(w|M_C) \tag{1.7}$$

Here, the basic ranking component is the (negated) cross-entropy $H(M_Q; M_D)$, which is normalized by the cross-entropy $H(M_Q; M_C)$.

Concluding, the probabilistic formulation of the prototypical IR-task: $P(L|D, Q)$ can be developed in two different ways; one starting from documents, the other one starting from queries. After applying Bayes' rule and transforming to log-odds, both variants can be rewritten to a sum of a likelihood ratio and the odds of the prior probability. The denominator in the likelihood ratio is a key element to ensure comparable scores of the events which are compared (document descriptions in the case of the document likelihood variant and query descriptions in the case of the query likelihood variant). Apart from the fact that the likelihoods of documents and queries have to be normalized in order to model $P(L|D, Q)$ (Bayes' rule), we have seen that we have to apply some corrections to account for differences in length, since the basic model is based on descriptions of similar length.

We summarize the length normalization aspects of the various models in table 1.1, the table lists 'yes' if the particular model inherently accounts for length differences and 'no' if an external length normalization is required.

Model name	query length normalization	document length normalization	reference
BIR	no	yes	Robertson & Sparck Jones
document likelihood ratio	yes	no	Lavrenko & Croft
BII	yes	no	Maron&Kuhns, Fuhr
query likelihood (ratio)	no	yes	Hiemstra, Miller et al.

Table 1.1. Length normalization of probabilistic IR models

2.2 A single model for ad-hoc and tracking?

From an abstract matching point of view, there are no major differences between the tracking and ad-hoc task. There is some text, which describes the domain of interest of the user, and subsequently a list of documents has to be ranked according to relevance to that description. So, it is a valid question to ask, whether we could define a model, which works well for both tasks. As we have argued, such a model has to be insensitive to length differences both for queries and for documents. Indeed, it seems valid to say that a good tracking system would work well for ad-hoc tasks as well, since the additional constraint concerning score normalization across topics does not affect the rank order of the documents.

However, when we compare the tasks in more detail, there are certainly many differences between the ad-hoc and the tracking task. First of all, the "matching" situation is extremely asymmetric for the ad-hoc task: a query is usually very short in comparison with a document. Moreover, not all words in the query are about the domain of interest, some serve to formulate the query. There are no phrases like "Relevant documents discuss X" in TDT topics. The tracking task does not provide any query at all, just one or more example stories. In that respect, "matching" is much more symmetric for tracking. The asymmetry of the ad-hoc task is probably the reason why the query likelihood approach is so successful: a document contains a much larger foothold of data to estimate a language model than a query (Lafferty and Zhai, 2001). This preference is probably not so clear-cut for tracking. Indeed, BBN has experimented with both directions and found that they complement each other (Jin et al., 1999). Also, relevance in TDT is different from relevance in the traditional ad-hoc task, since TDT is concerned with events. Although the tracking task lacks supervised relevance feedback, unsupervised feedback (topic model adaptation) is allowed. In a way, this procedure is related to pseudo-feedback techniques in IR. However, the tracking task lacks the notion of the "top-N" documents, i.e. unsupervised feedback has to be based on absolute instead of relative scores, which is certainly more complicated.

In our experiments, we do not want to rule out specific models a-priori on the basis of the differences between ad-hoc and tracking, but instead will investigate whether probabilistic language models which are successful for the ad-hoc task can be adapted for tracking. We will study the necessity and relative effectiveness of normalization procedures. Therefore we will test both directions of the generative model for the tracking task.

2.3 Ranking with a risk metric: KL divergence

Recently, Lafferty and Zhai proposed a document ranking method based on a risk minimization framework(Lafferty and Zhai, 2001). As a possible instantiation of this framework, they suggest to use the relative entropy of Kullback-Leibler divergence between a distribution representing the query and a distribution the document $\Delta(M_Q||M_D)$ as a loss function. The KL divergence is a measure for the difference between two probability distributions over the same event space.

$$\Delta(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \tag{1.8}$$

KL divergence has an intuitive interpretation, since the KL divergence is either zero when the probability distributions are identical or has a positive value, quantifying the difference between the distributions by the number of bits which are wasted by encoding events from the distribution P with a "code" based on distribution Q. However, KL also has some less attractive characteristics: it is not symmetric and does not satisfy the triangle inequality and thus is not a metric(Manning and Schütze, 1999).

The relationship between the KL divergence and language models for IR was initially discussed by Ng (Ng, 2000). The relationship of (1.7) with $\Delta(M_Q||M_D)$ is as follows:

$$\Delta(M_Q||M_D) = \sum_w P(w|M_Q)\log\frac{P(w|M_Q)}{P(w|M_D)} + \sum_w P(w|M_Q)\log\frac{P(w|M_C)}{P(w|M_C)} \quad (1.9)$$

after reformulation:

$$NLLR(Q;D,C) = \Delta(M_Q||M_C) - \Delta(M_Q||M_D) \quad (1.10)$$

It is tempting to interpret this equation as a subtraction of two values of a similar metric. However, this is invalid. Informally, we might interpret the generalized (or story length normalized) log likelihood ratio by taking a closer look at the two components: a score based on NLLR is high when $\Delta(M_Q||M_C)$ is high and $\Delta(M_Q||M_D)$ is low. This means that a story has a higher score when it contains specific terminology, i.e. is dissimilar from the background collection model and when its distribution is close to the topic distribution. For ad-hoc search, $\Delta(M_Q||M_D)$ is essentially equivalent to the length normalized query likelihood (1.6) since the query entropy $H(M_Q) = \sum_w P(w|M_Q)\log P(w|M_Q)$ is a constant which does not influence document ranking. Several authors have presented KL divergence as a valid and effective ranking model for ad-hoc IR tasks (Ogilvie and Callan, 2001; Lavrenko et al., 2002b). They consider query likelihoods as a derived form of the more general KL divergence. Since we are looking for a general model, which is useful for both the ad-hoc and the tracking task, we will evaluate the KL-divergence measure for tracking in addition to the models presented in Section 2.

2.4 Parameter estimation

In the previous sections, we have only marginally talked about how unigram language models can be estimated. A straightforward method is to use maximum likelihood estimates, but just like language models for speech recognition, these estimates have to be smoothed. One obvious reason is to avoid to assign zero probabilities for terms that do not occur in a document because the term probabilities are estimated using maximum likelihood estimation. If a single query term does not occur in a document, this would amount to a zero probability of generating the query. There are two ways to cope with this. One could either model the query formulation process with a mixture model based on a document model and a background model or assume that all document models can in principle generate all terms in the vocabulary, but that irrelevant terms are generated with a very small probability.

A simple yet effective smoothing procedure, which has been successfully applied for ad-hoc tasks in linear interpolation (Miller et al., 1999; Hiemstra, 1998). Recently other smoothing techniques (Dirichlet, absolute discounting) have been evaluated (Zhai and Lafferty, 2001). These authors argued that smoothing actually has two roles: i) improving the probability estimates of a document model, which is especially important for short documents, and ii) “facilitating” the generation of common terms (a *tfidf* like function). Dirichlet smoothing appears to be good for the former role and linear interpolation (which is also called Jelinek-Mercer smoothing) is a good strategy for the latter function. In the experiments reported here, we have smoothed all generating models by linear interpolation. We did some preliminary experiments with Dirichlet smoothing, but did not find significant improvements.

Linear interpolation based smoothing of e.g. a topic model is defined as follows:

$$P(w|M_T) \;=\; \lambda P(w|M_T) + (1-\lambda)P(w|M_C)) \tag{1.11}$$

The probability of sampling a term w from topic model M_T is estimated on the set of training stories for M_T using a maximum likelihood estimator. This estimate is interpolated with the marginal $P(w|M_C)$ which is computed on a large background corpus (the entire TDT2 corpus).

2.5 Parametric score normalization

We have seen in Section 2.1 that it is easy to normalize generative probabilities for differences in length. Length normalized generative probabilities have a sound Information Theoretic interpretation. Length might not be the only topic dependent score dependency we we have to correct for. For example in a model which is based on the query likelihood: $NLLR(Q; D, C)$ with smoothing based on linear interpolation, the median of the score distribution for each topic will differ, since it is directly correlated with the average specificity of topic terms. Let’s look at a couple of extreme cases of title queries from the TREC ad-hoc collection:

Query 403: osteoporosis A query of a single very specific word will yield document scores with high scores for those documents containing "osteoporosis". Since $P(w|M_D)$ is much higher than $P(w|M_C)$, the term weight is essentially determined by the ratio $\log(P(w|M_D)/P(w|M_C))$. Documents that do not contain the term "osteoporosis" do all have the constant score $\log((1-\lambda))$ due to smoothing.

Query 410: Schengen agreement This query consists of a quite specific proper name and the fairly general term "agreement". The contribution of "Schengen" to the total score of a document is much higher than "agreement". If a document does not contain "Schengen" it will not be relevant, therefore the score distributions between relevant documents are well separated[4].

Query 422: heroic acts This query does not contain any rare terms, consequently document scores of relevant documents are lower.

Even though maximum likelihood procedures normalize for most of the length variations in topics for $NLLR(S;T,C)$ models, we still expect length dependencies in the scores because the generating models are smoothed. A longer topic will have a higher probability to have overlapping terms with stories than a shorter topic, which we expect to see in the scores.

The examples make clear that the score distribution of relevant documents (say the documents that contain most of the important terms) is dependent on the query. Queries formulated with mostly specific terms, will produce higher scores. The score distribution of non-relevant documents containing any of the query terms does also depend on the query. A perfect tracking system would produce separated distributions of relevant and non-relevant stories with equal medians and variances across topics, because of the single threshold. In reality, distributions are never perfectly separated (i.e. the situation of $Precision = Recall = 1$). But we might be able to normalize score distributions.

Score distributions have been studied by different researchers in the context of collection fusion (Baumgarten, 1997; Baumgarten, 1999; Manmatha et al., 2001) or adaptive filtering (Arampatzis and Hameren, 2001). These researchers tried to model score distributions of relevant and non relevant documents by fitting the observed data with parametric mixture models (e.g. Gaussian for relevant documents and exponential or Gamma for non relevant documents). If the parametric models are a good fit of the data, it just suffices to estimate the model parameters to calculate the probability of relevance at each point in the mixture distribution. Unfortunately, we have very little training data for the distribution of the relevant documents in the case of tracking, so an approach like (Manmatha et al., 2001) is not feasible here. Instead, we could try to just estimate the parameters of the model for the non-relevant stories and assume that the concentration of relevant documents in the right tail of this distribution is high and hope that there is a more or less similar inverse relationship between the density of non-relevant and relevant stories in this area of the curve. This normalization strategy was proposed and evaluated for TDT tasks by researchers at BBN (Jin et al., 1999). They modeled the distribution of non relevant documents by a Gaussian distribution, which can be justified by the central limit theorem for some of the models we have discussed. Indeed, the topic likelihood model score can be seen as a sum of independent random discrete variables. When a topic is long enough, the distribution can be approximated by the Gaussian distribution. It is unclear, whether this also holds for the story likelihood model, since the score is composed of a different number of variables for each story.

We implemented the Gaussian score normalization as follows: For each topic we calculated the scores of 5000 stories taken from the TDT Pilot corpus, we assumed these were non-relevant, since they predate the test topics[5]. We subsequently computed the mean and standard deviation of this set of scores. These distribution parameters were used to normalize the raw score τ in the following way:

$$\tau' = (\tau - \mu)/\sigma \tag{1.12}$$

3. Experiments

The generative models presented in the previous section will now be compared on two different test collections. Before presenting the actual data, the models will be briefly re-presented in Section 3.1 followed by background information about the test collections and test metrics that we used in Sections 3.2 and 3.3.

3.1 Experimental conditions

For our tracking experiments we plan to compare the following models:

Normalized Story likelihood ratio:$NLLR(S;T,C)$ This is the model described in (1.3), which can also be seen as a normalized cross-entropy.

Normalized Topic likelihood ratio:$NLLR(T;S,C)$ This is the model described in (1.7), also a normalized cross-entropy.

KL divergence: $\Delta(S||T)$ and $\Delta(T||S)$ Recently, several researchers have argued that the Kullback-Leibler divergence can be viewed as a general model underlying generative probabilistic models for IR.

The first two models are motivated by the probability ranking principle. Query likelihood ratio is based on a model for ranking queries, but can be used to rank documents. The KL divergence model is motivated as a loss function in a risk minimization framework, which does not explicitly model relevance.

Apart from comparing the effectiveness of the models as such, we will investigate the relative importance of several normalization components that are inherent to the models, for example the length normalization and the fact that the first two models compare entropy with respect to a common ground. We also evaluate the effectiveness of the Gaussian normalization and its interaction with different smoothing techniques.

3.2 The TDT evaluation method: DET curves

The TDT community has developed its own evaluation methodology. Because some of the plots further on in this article show results that were produced by this method, it is necessary to familiarize the reader with some of its details. All of the TDT tasks are cast as detection tasks. In contrast to TREC experiments, the complete test set for each topic of interest is annotated for relevance. Tracking performance is characterized in terms of the probability of miss and false alarm errors ($P_{Miss} = P(\neg ret|target)$ and $P_{FA} = P(ret|\neg target)$). To speak in terms of the more established and well-known precision and recall measures: a low P_{Miss} corresponds to high recall, while a low P_{FA} corresponds to high precision. These error probabilities are combined into a single cost measure C_{Det}, by assigning costs to miss and false alarm errors (Doddington and Fiscus, 2002):

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{\neg target} \tag{1.13}$$

where C_{Miss} and C_{FA} are the costs of a miss and a false alarm respectively; P_{Miss} and P_{FA} are the conditional probabilities of a miss and a false alarm respectively; P_{target} and $P_{\neg target}$ are the a priori target probabilities ($P_{\neg target} = 1 - P_{target}$).

Then C_{Det} is normalized so that $(C_{Det})_{Norm}$ can be no less than one without extracting information from the source data:

$$(C_{Det})_{Norm} = \frac{C_{Det}}{min(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{\neg target})} \tag{1.14}$$

Thus the absolute value of $(C_{Det})_{Norm}$ is a direct measure of the relative cost of the TDT system (Doddington and Fiscus, 2002).

The error probability is estimated by accumulating errors separately for each topic and by taking the average of the error probabilities over topics, with equal weight assigned to each topic. The following parameters were determined a-priori: $C_{Miss} = 1$, $C_{FA} = 0.1$, and $P_{target} = 0.02$. The Detection Error Tradeoff (DET) curve is the equivalent of a precision-recall plot for ad-hoc experiments. The DET plot shows what happens when the decision threshold of the tracking system performs a sweep from an (infinitely) high value to an (infinitely) low value. Obviously, at the beginning of the parameter sweep, the system will have zero false alarms but will not detect any relevant stories either and moves to the opposite end of the trade-off spectrum when the threshold is decreased. An example DET plot is Figure 1.1. A good curve in a DET plot is a relatively straight curve with a negative slope. The steeper the curve, the better.

We can prove that there is a simple relationship between the derivative of the DET curve (the slope of a tangent line at each point in the curve) and the probability of relevance. First we define P_{Miss} and P_{FA} as a function of the probability of relevance at rank n^6:

$$P_{Miss}(n) = \frac{\int_1^N P_r(n) - \int_1^n P_r(n)}{\int_1^N P_r(n)} \tag{1.15}$$

$$P_{FA}(n) = \frac{\int_1^n (1 - P_r(n))}{\int_1^N (1 - P_r(n))} \tag{1.16}$$

In these equations, $P_r(n)$ is the probability of relevance at rank n, N is the total number of stories and R is the total number of relevant stories. Now the slope of the DET curve as a function of n can be defined as:

$$\frac{\partial P_{Miss}}{\partial n} / \frac{\partial P_{FA}}{\partial n} = \frac{-P_r(n)}{1 - P_r(n)} \cdot \frac{\int_1^N (1 - P_r(n))}{\int_1^N P_r(n)} = \frac{-P_r(n)}{1 - P_r(n)} \cdot \frac{N - R}{R} \tag{1.17}$$

Formula (1.17) makes it easier to interpret the shape of a DET curve. A straight line means that the probability of relevance is constant for all ranks. An example of such a curve is the system that assigns scores to stories in a random fashion. If the curve is convex at some point, this means that the probability of relevance is decreasing with rank n. If the DET curve is concave at a certain point, the probability of relevance is increasing with n. A good normalized system would have a high relatively constant probability of relevance at the top ranks , followed by a short section where the probability of relevance gradually drops to a lower level and then a low relatively constant probability of relevance for the lower ranks. This would yield an extremely convex DET curve, consisting of two almost straight lines, not far from the axes of the plot and connected by a round convex segment where the probability of relevance drops from a high to a low value.

Note that the DET curves produced by the TDT evaluation software have custom scales - partly linear and partly logarithmic - in order to magnify certain areas of the curve. This has the effect that straight descending curves become concave in the logarithmic parts of the graph. This effect is doubled in the double-logarithmic part of the graph (the lower left corner). Straight curves are not transformed in the linear part of the graph (upper right part). The probability of relevance based definition of slope (1.17) is thus only valid for the linear - linear part of the graph. Still, convexness (or the absence of concave segments) indicates that the system produces well normalized score distributions.

3.3 Description of the test collections

Currently, the Linguistic Data Consortium (LDC) has three corpora available to support TDT research[7] (Cieri et al., 2000). The TDT-Pilot corpus contains newswire and transcripts of news broadcasts, all in English, and is annotated for 25 news events. The TDT2 and TDT3 corpora are multilingual (TDT2: Chinese and English, TDT3: Chinese, English, and Arabic) and contain both audio and text. ASR transcriptions and close captions of the audio data as well as automatic translations of the non-English data are also provided. TDT2 and TDT3 are annotated for 100 and 120 news events respectively.

We conducted several experiments on a subset of the TDT2 corpus to investigate the effect of stemming, the value of the smoothing parameter λ and a comparison of the two different orientations of generative model for tracking: generating the topic or the story. These experiments are reported in (Spitters and Kraaij, 2001). In this paper we describe experiments focused on score normalization. These experiments were conducted using the Jan-Apr part of the TDT2 corpus as training and development data, and the May-June part as the evaluation data (17 topics). Our study is limited to a simplified dataset, we work with the output of automatic speech recognizers, which is pre-segmented. The

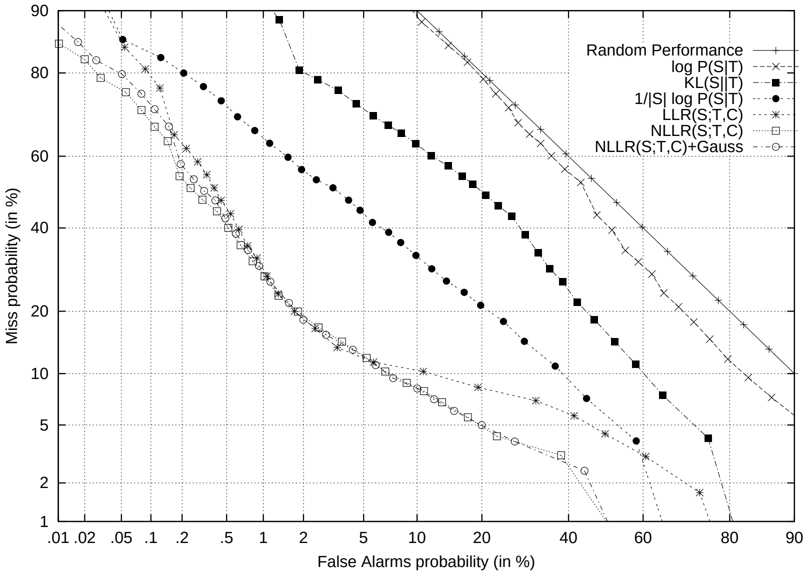

Figure 1.1. Comparison of different tracking models based on $P(S|T)$.

foreign language material has been processed by a Machine Translation system. We will not study source specific dependencies, i.e. we regard the dataset as a uniform and monolingual collection of news stories. All experiments were done with just one training story per topic.

Because experimentation with tracking is a time consuming process, we also simulated a tracking task by using TREC ad-hoc runs. We replicated an experiment presented by Ng (Ng, 2000) who simulated a binary classification task on TREC ad-hoc data with a fixed threshold. We will discuss further details in Section 3.5.

3.4 Experiments on TDT test collection

We will first present a comparison of the basic models which have $P(S|T)$ as their core element: topic likelihood models. All experiments are based on smoothing by linear interpolation with a fixed $\lambda = 0.85$.

Figure 1.1 shows the results of several variant models in a DET curve. The basic story likelihood model $P(S|T)$ is hardly better than a random system (with a constant $P_r(n)$). This is not surprising, since the likelihood is not normalized. The relative effect of the two normalization components i.e. normalizing by the a-priori story likelihood $NLLR(S;T,C)$ and story length normalization is quite different. Taking the likelihood ratio is the fundamental step, which realizes the *idf*-like term weighting and converts likelihood ranking to log-odds ranking (cf. formula (1.2)). Story length normalization removes some variance in the scores due to length differences and improves upon the LLR model for most threshold levels. Our basic tracking model (NLLR) combines both normalization steps.

Surprisingly, the performance of $\Delta(S||T)$ is even worse than the length normalized likelihood $H(S;T)$. The Kullback-Leibler divergence can be seen as an entropy normalized version of the latter: $\Delta(S||T) = -H(S;T) + H(S)$, whereas the (length) normalized log likelihood ratio normalizes by the cross-entropy with the background collection: $NLLR(S;T,C) = -H(S;T) + H(S;C)$. Our experimental results make clear that normalizing with entropy deteriorates results, whereas normalizing with $P(S|C)$ (or its length normalized version $H(S;C)$) is an essential step in achieving good results.

We repeated the same experiments for the reversed orientation: generating the topics from the stories. Results are plotted in Figure 1.2. The relative performance of the $P(T|S)$ based variant models is roughly equivalent to the variants of the $P(S|T)$ models with the exception of the models, which are not based on a likelihood ratio. Again, the main performance improvement is achieved by normalizing $P(T|S)$ with the prior likelihood $P(T|C)$, which is equivalent to ranking by log-odds of being liked. Length normalization improves performance at both ends of the DET-plot and results in a straighter curve. The length normalized likelihood model $1/|T| \log P(T|S)$ performs worse than its reverse counterpart. This is due to the fact that scores are not normalized for average term specificity across topics. An even more striking phenomenon is the step-like behaviour of the unnormalized $P(T|S)$. This is due to the fact that the score distributions of plain $P(T|S)$ are linearly dependent on topic lengths and consequently their medians are located quite far apart. We will illustrate this effect by some boxplots.

A boxplot is a graphical summary of a distribution, showing its median, dispersion and skewness. Therefore boxplots are extremely helpful to compare many different distributions. A boxplot is defined by five datapoints: the smallest value, the first quartile, the median, the third quartile and the largest value. The area between the first and third quartile (interquartile range) is depicted by a box, with a line marking the median. (Figure 1.5 is a good example.) The boxplots in this paper also have whiskers that mark either the smallest or largest value or the

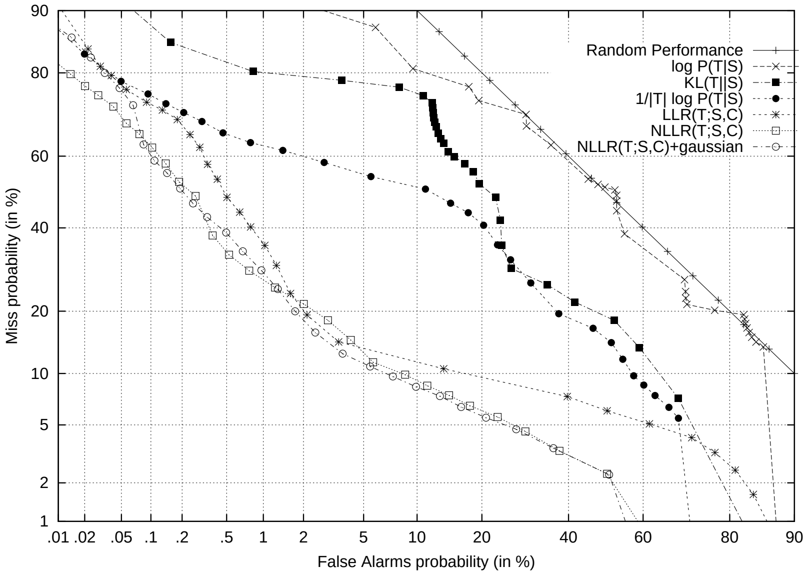

Figure 1.2. Comparison of different tracking models based on $P(T|S)$.

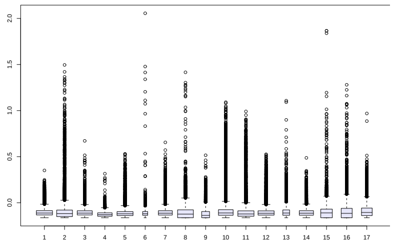

Figure 1.3. Score distributions of $NLLR(T; S, C)$

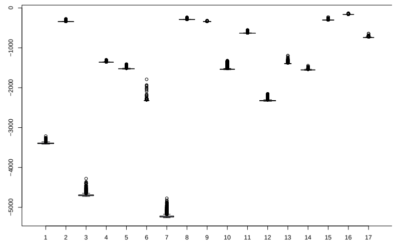

Figure 1.4. Score distributions of $P(T|S)$

area that extends 1.5 times the interquartile range from the first or third quartile.

Figures 1.3 and 1.4 show boxplots of the distributions of $NLLR(T; S, C)$ and $P(T|S)$ respectively. The first plot shows that the bodies of the distributions for all topics are quite well aligned. The distributions are skewed and have a long right tail, because they are in fact mixtures of a large distribution of relevant stories and a small distribution of non-relevant stories with a higher median. Figure 1.4 gives an explanation why the DET plot curve of this model is so wobbly: the distributions of the individual topics do not even overlap in a few cases: lowering the threshold will bring in the stories of each topic as separate blocks. This means that the probability of relevance will increase and decrease locally as we decrease the threshold, causing convex and concave segments (cf. Section 3.2). Because the boxes are hardly visible in both cases, we show an example of a more dispersed distribution: $KL(T||S)$ in Figure 1.5. The fact that the distributions lack a long right tail is a sign that relevant and non-relevant documents are probably not well separated. Finally, an example of well-aligned symmetrical distributions is $LLR(S; T, C)$ in Figure 1.6. The symmetry is due to the fact that scores are not length normalized, long stories that do not have word overlap with the topic will have high negative scores, long stories with good word overlap with the topic will have high positive scores.

Figure 1.7 shows that indeed there is some topic length effect for the $NLLR(S; T, C)$ as we hypothesized in Section 2.5. For example, the first topic has length 395 and the second has length 43, which results in lower scores for the bulk of the distribution. Figure 1.8 shows score distributions of the same model after applying Gaussian normalization. Indeed the boxes are better aligned, but differences are small. The normalization resulted however in some performance loss in the high precision

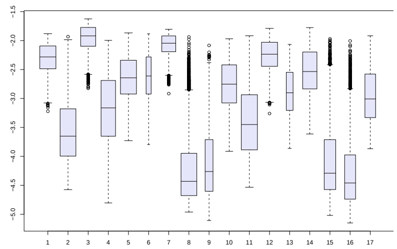

Figure 1.5. $KL(T||S)$

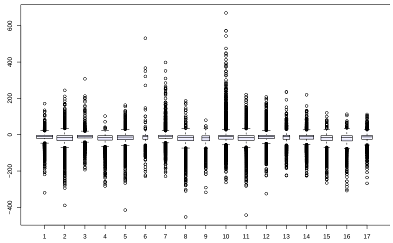

Figure 1.6. $NLLR(S|T)$

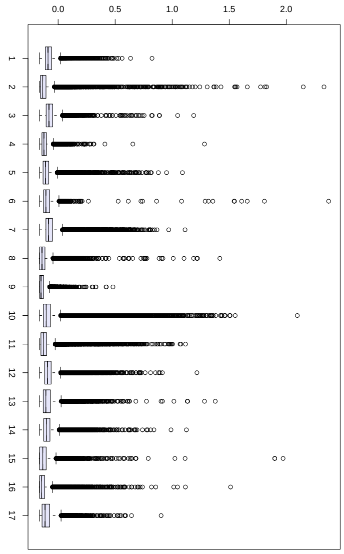

Figure 1.7. $NLLR(S;T,C)$

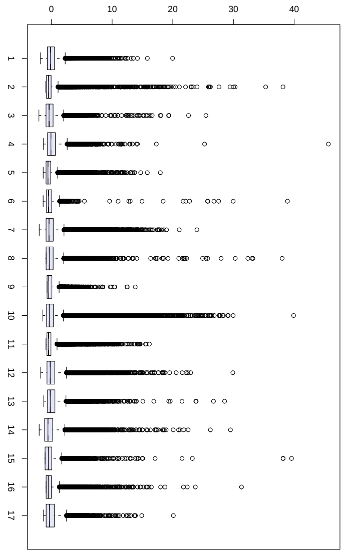

Figure 1.8. $NLLR(S;T,C)$ + Gaussian normalization

area, cf. Figure 1.1. We have also applied Gaussian normalization to the $LLR(S;T,C)$ model, which is not normalized for story length. In this case, the Gaussian normalization deteriorated results, even though medians were well aligned. We think that this is due to the fact that the variance in the score distribution is due to differences in length, which can be normalized in a more effective way. Gaussian normalization of the model in the reverse orientation: $NLLR(T;S,C)$ had similar effects: a small performance loss in the high precision area and for the rest roughly equivalent to the not Gaussian normalized version (cf. Figure 1.2). Further investigation is needed in order to understand why the Gaussian normalization is not effective. There are several possibilities: i) scores are already quite well normalized, ii) the score distribution differs too much from the normal distribution, or iii) outliers hurt the estimation of the distribution parameters.

Since both orientations of the NLLR model work well, there might be some potential to improve results by a combination of the scores of both models. We did some initial experiments which were based on a simple score averaging procedure. A side effect of this method is that scores become symmetric. It is exactly this symmetrical NLLR model that had proven to be effective for the TDT detection task (Spitters and Kraaij, 2002). The resulting system performed worse than each of the components, but after applying Gaussian normalization the system was a little bit more effective than a model based on just a single orientation. Further research is needed to find an optimal combination/normalization procedure.

3.5 Simulating tracking on TREC Ad-Hoc data

We complemented the runs on the TDT2 corpus with experiments on TREC ad-hoc data. The main reason is that most data was available already, and provided a rich resource for research on score normalization. Since ad-hoc runs output a list of scored documents, we could simulate a $NLLR(Q;D;C)$ tracking system, by placing a threshold. We applied two methods to implement this idea. The first method is based on `trec_eval`, the second on the TDT evaluation software.

The basic idea is to evaluate all 50 topics of an ad-hoc run by a single threshold. Standard `trec_eval` does not support this kind of evaluation. However, it can be simulated by replacing all topic-id's in both the runs and the qrel file by a single topic-id. Of course, this evaluation is different from TDT eval, since this method does not involve topic averaging, so topics with many relevant documents will dominate the results. Still, this evaluation is a quick and easy method to assess score stability across topics when TDT evaluation software is not available. We tested this method on the TREC-8 ad-hoc test collection, for both title and full queries.

run name	title (tracking)	title (ad-hoc)	full (tracking)	full (ad-hoc)
$P(Q\|D)$	0.0874	0.2322	0.1358	0.2724
$LLR(Q;D,C)$	0.1334	0.2321	0.1581	0.2723
$NLLR(Q;D,C)$	0.1294	0.2324	0.1577	0.2723
$\Delta(Q\|\|D)$	0.0845	0.2322	0.1356	0.2723

Table 1.2. Tracking simulation on TREC-8 Ad-Hoc collection (mean average precision)

Table 1.2 shows the results of our experiments, using four weighting schemes: straight (log) query likelihood, log-likelihood ratio, normalized log-likelihood ratio and Kullback-Leibler . We see that the influence of the particular normalization strategy is quite strong on the tracking task, while - as was expected - there is no influence on the ad-hoc task. Indeed the normalization strategies just add topic specific constants, which do not influence the ad-hoc results. There seems to be no big difference between LLR and NLLR, but that might be due to the averaging strategy, which is not weighted across topics. NLLR is a bit less effective than LLR for title queries, but that can be explained by the difference in query term specificity for short (1-3 word) queries. A single word TREC title query must be very specific (e.g. topic 403: "osteoporosis") in order to be effective. Two and three word queries often use less specific words and thus their scores will be lower in the NLLR case, which is normalized for query length. Still two or three word queries can be just as effective as one word queries, so there is no reason to down-normalize their scores. This effect was confirmed by the boxplots for these runs, shown in Figures 1.9 and 1.10. The title queries with the highest NLLR scores (403 and 424) are single word queries. The boxplots show a mix of topics to visualize the topic normalization, the score distributions of the first 25 topics (topic 401-425) are based on title queries, the rightmost 25 distributions are based on the full queries (topic 426-450).

The Kullback Leibler divergence based run really performs disappointingly. We can conclude that KL as such is not a suitable model for tracking, at least not for models estimated with maximum likelihood estimation.

We also ran the TDT evaluation scripts on the TREC data after applying a conversion step. The difference with the previous method, is that

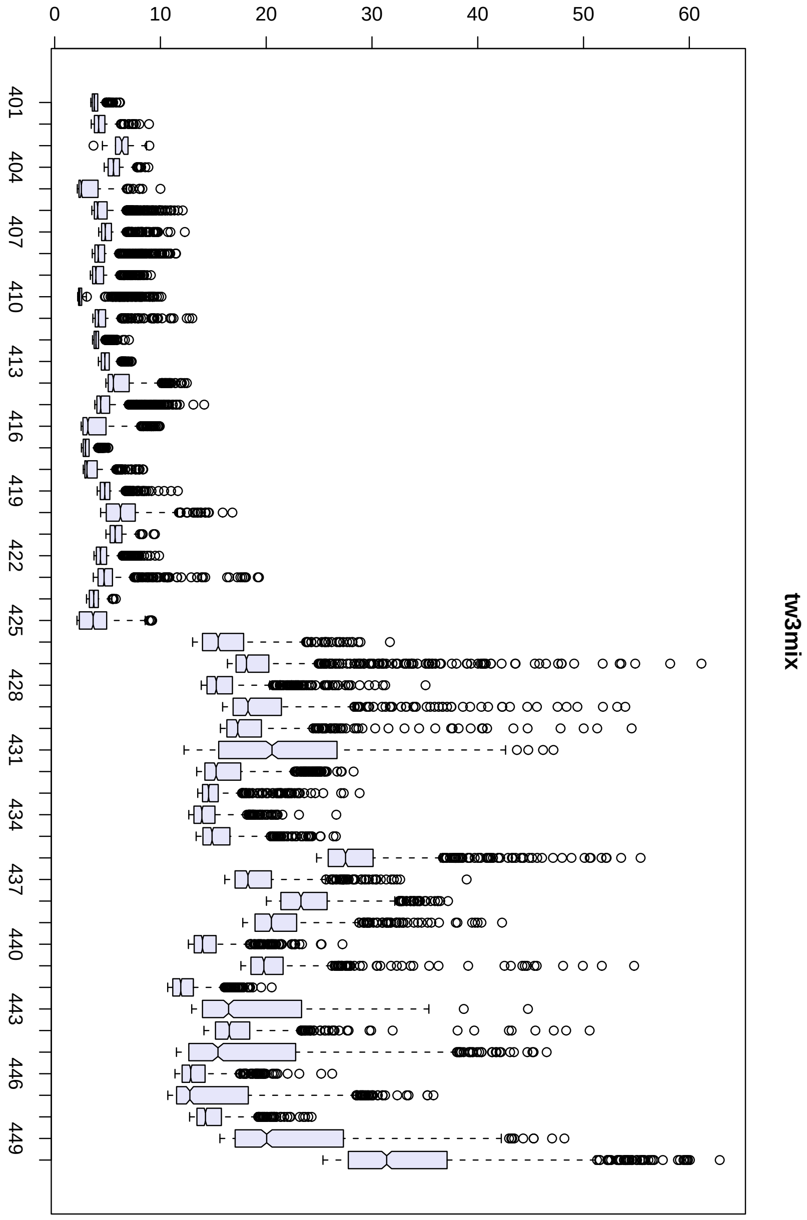

Figure 1.9. Tracking simulation: LLR

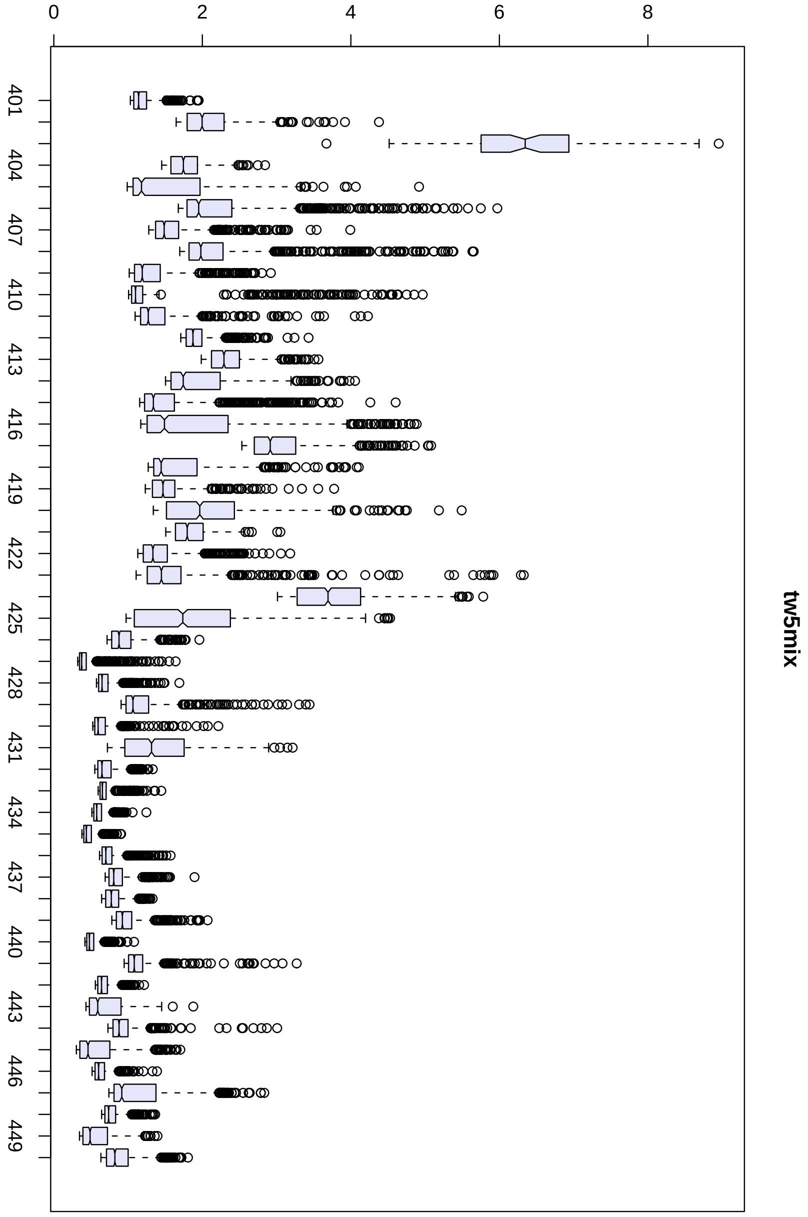

Figure 1.10. Tracking simulation: NLLR

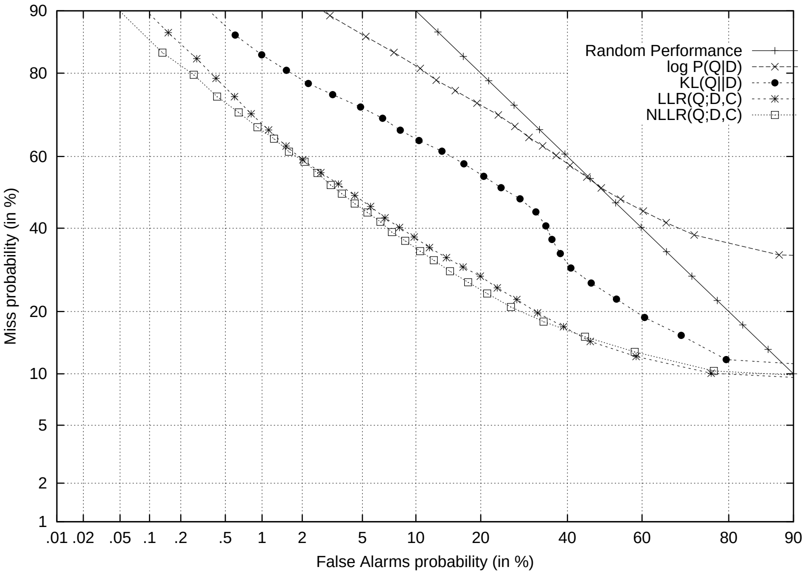

Figure 1.11. DET plot of tracking simulation on the TREC 8 ad-hoc full topics

the TDT evaluation procedure averages P_{FA} and P_{Miss} across topics. The results of the run based on the full topics are shown in plot 1.11. The best performance is reached by NLLR, which is just a bit better than LLR. Again KL yields a very disappointing result.

4. Discussion

One of the main challenges of designing a tracking system is to normalize scores across topics. Since topics are of a very different nature, and there is no direct relationship between the score distribution of the models and probability of relevance, this is quite a hard task. An ideal system would produce the probability of relevance of each test story/document as a score. A system could then be optimized for a certain cost or utility function, by setting a threshold on a certain probability of relevance value. However, there is only an indirect relationship between score distribution and probability of relevance. We have seen that scores can be dependent on story and/or topic length and on the term specificity of their formulation. Some topics are "easy" i.e. the score distributions of relevant and irrelevant stories are well separated. We have tried to cope with these differences using different techniques, i) we used a model with inherent normalization: the (normalized) log likelihood ratio ii) we tried to model the score distributions themselves.

The log-likelihood ratio based tracking models are directly derived from the probability of relevance and thus have the advantage that the scores have a clear interpretation and a common reference point. They compare the generative probability of a story given the topic (or vice versa) in comparison with the a-priori probability. Likelihood ratios are in fact a form of statistical hypothesis tests, where each generative model is one hypothesis. We have previously reported that the results of our NLLR based system for the official TDT 2000 tracking task were competitive (Spitters and Kraaij, 2001). We therefore conclude that language models can form the basis for an effective tracking system indeed, provided that the models are properly normalized. Our experiments with TDT and TREC data showed that normalizing by an a-priori model is a key point. But the function of the normalizing likelihood in the denominator is different for the two orientations of the model. In the story likelihood case, the $P(S|C)$ component normalizes scores for across story differences in term specificity, whereas in the topic likelihood case, the $P(T|C)$ component normalizes scores for across topic differences in term specificity. Scores can be normalized further by applying length normalization. Both orientations of the model have comparable tracking effectiveness for the case of a single training document.

We also evaluated methods for score normalization which try to fit the distribution of non-relevant stories by a Gaussian distribution. This normalization was not really effective for the $NLLR(S;T,C)$ model and even seriously hurt the effectiveness of the $LLR(S;T,C)$ model. We think that the $LLR(S;T,C)$ model is not suitable for Gaussian normalization since the score variance is dominated by differences in story length, which should be removed prior to Gaussian normalization. BBN has reported favourable results with Gaussian normalization. We conjecture that Gaussian normalization could work for their **IR** model, which is equivalent to $P(T|S)$; the straight topic likelihood. Gaussian normalization is able to normalize across topic differences. However, a simpler method is to work with the likelihood ratio $P(T|S)/P(S)$ instead. After all, unlike ad-hoc the denominator $P(S)$ is not a constant.

Despite the intuitive appeal of KL - measuring the dissimilarity between distributions - our experiments with KL for tracking yielded disappointing results for both orientations $\Delta(S||T)$ and $\Delta(T||S)$. The Kullback-Leibler divergence has usually been proposed in an ad-hoc query-likelihood context. In that case KL reduces to pure query-likelihood, since the normalizing entropy $H(Q)$ in the KL divergence can be discarded because it is a constant. This cannot be done in a tracking context and we have seen that normalizing a cross entropy by its entropy is not effective in a tracking context. We have also shown that normalizing by the prior probability of a topic as measured by its cross entropy with the collection model is effective. Informally we could say that the KL divergence based scores are not properly normalized. The problem of using KL divergence for tracking is that the scores lack a common point of reference. Dissimilarity is measured on the basis of different models and since KL is not a metric, these scores cannot be compared. A more formal criticism on the use of KL divergence for tracking is that KL based models lack the notion of relevance. We have seen that both orientations of the normalized log likelihood ratio, which are direct derivations of probability of relevance based ranking are effective for tracking.

This analysis has been recently confirmed by independent research in the area of story link detection (Lavrenko et al., 2002a). Lavrenko found that pure KL is not effective for story link detection and proposed the so-called "Clarity-adjusted KL" topic similarity measure to correct for the fact that KL does not concentrate on informative (in the *idf* sense) terms when computing the (dis)similarity score. This adjusted KL measure is defined as $-\Delta(T||S) + Clarity(T)$, where clarity is defined as $\Delta(T||C)$ (Cronen-Townsend et al., 2002). Indeed, when comparing this definition to formula (1.10), the Clarity-adjusted KL divergence seems to be equivalent[8] to the normalized log-likelihood ratio. The NLLR similarity measure can thus be motivated by two frameworks: i) direct derivation from the log-odds of relevance ii) clarity adjusted version of KL divergence.

We also evaluated the query-likelihood models by a simulation of the tracking task on the TREC-8 ad-hoc collection. We know that there is a real difference between ad-hoc topics and TDT topics, this difference is one of the reasons that score normalization effectiveness differs across these tasks. TDT topics are just stories describing a particular event. Ad-hoc topics are structured queries which are stated in a particular jargon. The bag-of-words approach we took for ad-hoc query construction showed a clearly visible difference between title queries and full queries. Even our best normalization strategy (NLLR) could not "smooth out" the differences in score distributions between these two types of queries. We plan to develop topic-type (e.g. title versus full) specific query distribution estimation methods, which we hope will enable us to further normalize scores.

5. Conclusions

Our aim was to find generative probabilistic models that work well both for the ad-hoc task and the tracking task, because we realized that a tracking system puts just one additional constraint on matching function: across topic comparability of scores. With the probability ranking principle as a starting point, we reviewed two lines of probabilistic modeling, either based on the document likelihood ratio or the query likelihood ratio. We evaluated variants of both models, based on length normalization and Gaussian normalization. We found that both orientations of the log-likelihood ratio work well. The essential normalization component in the NLLR model is the a-priori likelihood (or cross entropy) of the generated text in the denominator. Effectiveness can be further enhanced by length normalization.

We have not been able to show performance increase by Gaussian normalization. The NLLR model is related to the negated KL divergence since both measures are based on the cross entropy. We found that KL divergence is not an effective scoring function for tracking, because the scores are not comparable across topics (for $\Delta(T||S)$) or across stories (for $\Delta(S||T)$). The principal reason seems to be the fact that the application of KL divergence as a similarity measure for the tracking task lacks normalization with respect to a common reference distribution.

Acknowledgements

We would like to thank Kenney Ng for providing details about his normalization experiments with ad-hoc data and Jon Fiscus for implementing the TREC2TDT conversion script.

Notes

1. The logarithm converts products to summations, working with the odds results in a simple likelihood ratio after applying Bayes' rule.
2. This normalization is in fact the probabilistic justification of *idf* weighting
3. Matching scores are already length normalized in the language model of Ponte and Croft, since queries are represented as a binary vector defined on the complete collection vocabulary (Ponte and Croft, 1998).
4. This can clearly be seen in Figure 1.10, which we will discuss later.
5. However, some of these stories could be considered relevant under a more liberal definition. Removal of these outliers has been reported to improve parameter estimation(Jin et al., 1999)
6. The DET curve as a function of the rank is identical to the DET curve as a function of the score. The story with the highest score gets rank 1.
7. http://www.ldc.upenn.edu/Projects/TDT
8. Note however that in Lavrenko's framework, topic models are estimated using the relevance model technique.

References

Arampatzis, A. and Hameren, A. (2001). The score-distributional threshold optimization for adaptive binary classification tasks. In (Croft et al., 2001b), pages 285–293.

Baumgarten, C. (1997). A probabilistic model for distributed information retrieval. In Belkin, N. J., Narasimhalu, A. D., and Willet, P., editors, *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, pages 258–266. ACM Press.

Baumgarten, C. (1999). A probabilistic solution to the selection and fusion problem in distributed information retrieval. In (Hearst et al., 1999), pages 246–253.

Beaulieu, M., Baeza-Yates, R., Myaeng, S. H., and Järvelin, K., editors (2002). *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002).* ACM Press.

Cieri, C., Graff, D., Liberman, M., Martey, N., and Strassel, S. (2000). Large multilingual broadcast news corpora for cooperative research in topic detection and tracking: The TDT2 and TDT3 corpus efforts. *Proceedings of the Language Resources and Evaluation Conference (LREC2000).*

Croft, B., Callan, J., and Lafferty, J. (2001a). Workshop on language modeling and information retrieval. *SIGIR FORUM*, 35(1).

Croft, W., Harper, D., D.H.Kraft, and Zobel, J., editors (2001b). *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001).* ACM Press.

Cronen-Townsend, S., Zhou, Y., and Croft, W. (2002). Predicting query performance. In (Beaulieu et al., 2002).

Doddington, G. and Fiscus, J. (2002). The 2002 topic detection and tracking (TDT2002) task definition and evaluation plan. Technical Report v. 1.1, National Institute of Standards and Technology.

Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, 35(3):233–245.

Hearst, M., Gey, F., and Tong, R., editors (1999). *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99).* ACM Press.

Hiemstra, D. (1998). A linguistically motivated probabilistic model of information retrieval. In Nicolaou, C. and Stephanides, C., editors, *Research and Advanced Technology for Digital Libraries - Second European Conference, ECDL'98, Proceedings*, number 1513 in Lecture Notes in Computer Science. Springer Verlag.

Hiemstra, D. and Kraaij, W. (1999). Twenty-one at TREC-7: Ad hoc and cross language track. In Voorhees, E. M. and Harman, D. K., editors, *The Seventh Text REtrieval Conference (TREC-7)*, volume 7. National Institute of Standards and Technology, NIST. NIST Special Publication 500-242.

Jin, H., Schwartz, R., Sista, S., and Walls, F. (1999). Topic tracking for radio, tv broadcast and newswire. In *Proceedings of the DARPA Broadcast News Workshop.*

Kraaij, W., Pohlmann, R., and Hiemstra, D. (2000). Twenty-one at TREC-8: using language technology for information retrieval. In (Voorhees and Harman, 2000). NIST Special Publication 500-246.

Kraaij, W., Westerveld, T., and Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In (Beaulieu et al., 2002).

Lafferty, J. and Zhai, C. (2001). Probabilistic IR models based on document and query generation. In Callan, J., Croft, B., and Lafferty, J., editors, *Proceedings of the workshop on Language Modeling and Information Retrieval.*

Lavrenko, V., adn Edward DeGuzman, J. A., LaFlamme, D., Pollard, V., and Thomas, S. (2002a). Relevance models for topic detection and tracking. In *Proceedings of HLT 2002.*

Lavrenko, V., Choquette, M., and Croft, W. (2002b). Cross-lingual relevance models. In (Beaulieu et al., 2002).

Lavrenko, V. and Croft, W. (2001). Relevance-based language models. In (Croft et al., 2001b).

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in informatiion retrieval. In C.Nédellec and Rouveirol, C., editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 4–15.

Manmatha, R., Rath, T., and Feng, F. (2001). Modelling score distributions for combining the outputs of search engines. In (Croft et al., 2001b), pages 267–275.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT press.

Maron, M. and Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7:216–244.

Miller, D. R. H., Leek, T., and Schwartz, R. M. (1999). A hidden markov model information retrieval system. In (Hearst et al., 1999), pages 214–221.

Ng, K. (2000). A maximum likelihood ratio information retrieval model. In (Voorhees and Harman, 2000). NIST Special Publication 500-246.

Ogilvie, P. and Callan, J. (2001). Experiments using the lemur toolkit. In (Voorhees and Harman, 2001).

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In Croft, W., Moffat, A., van Rijsbergen, C., Wilkinson, R., and Zobel, J., editors, *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 275–281. ACM Press.

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

Robertson, S. and Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146.

Sparck Jones, K., Walker, S., and Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments. *ipm*, 36(6).

Spitters, M. and Kraaij, W. (2001). Using language models for tracking events of interest over time. In *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR2001)*.

Spitters, M. and Kraaij, W. (2002). Unsupervised event clustering in multilingual news streams. *Proceedings of the LREC2002 Workshop on Event Modeling for Multilingual Document Linking*, pages 42–46.

Voorhees, E. M. and Harman, D. K., editors (2000). *The Eigth Text REtrieval Conference (TREC-8)*, volume 8. National Institute of Standards and Technology, NIST. NIST Special Publication 500-246.

Voorhees, E. M. and Harman, D. K., editors (2001). *The Tenth Text REtrieval Conference (TREC-2001), notebook*, volume 10. National Institute of Standards and Technology, NIST.

Wayne, C. (2000). Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. *Proceedings of the Language Resources and Evaluation Conference (LREC2000)*, pages 1487–1494.

Zhai, C. and Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In (Croft et al., 2001b).