Probabilistic Aspects in Spoken Document Retrieval

Wolfgang Macherey, Hans Jörg Viechtbauer, and Hermann Ney

Abstract— **Accessing information in multimedia databases encompasses a wide range of applications in which spoken document retrieval (SDR) plays an important role. In SDR, a set of automatically transcribed speech documents constitutes the files for retrieval, to which a user may address a request in natural language. This article deals with two probabilistic aspects in SDR. The first part investigates the effect of recognition errors on retrieval performance and inquires the question, why recognition errors have only a little effect on the retrieval performance. In the second part, we present a new probabilistic approach to SDR that is based on interpolations between document representations. Experiments performed on the TREC-7 and TREC-8 SDR task show comparable or even better results for the new proposed method than other advanced heuristic and probabilistic retrieval metrics.**

Keywords— **spoken document retrieval, error analysis, probabilistic retrieval metrics**

Contents

I. Introduction

RETRIEVING information in large, unstructured databases is one of the most important tasks computers are used for today. While in the past, information retrieval focused on searching written texts only, the field of applications has since then extended to multimedia data, such as audio and video documents which are growing every day in broadcast and media. Nowadays, radio and TV stations hold huge archives containing numberless documents that were produced and collected over the years. However, since these documents are usually neither indexed nor cataloguized, the respective document collections are effectively not usable and thus, the data stocks are idle. Therefore, the need of efficient methods enabling content-based access to little or even un-structured multimedia archives is of eminent importance.

A. Spoken document retrieval

A particular application in the domain of information retrieval is the content based access to audio data in which *spoken document retrieval* (SDR) plays an important role. SDR extends the techniques developed in text retrieval to audio documents containing speech. To this purpose, the audio documents are automatically segmented and transcribed by a speech recognizer in advance. The resulting transcriptions are indexed and subsequently stored in large databases, thus constituting the files for retrieval, to which a user may address a request in natural language.

Over the past years, research shifted from pure text retrieval to SDR. However, since also state-of-the-art speech recognizers are still error-prone and thus far from perfect recognition, automatically generated transcriptions are often flawed, and not seldom they achieve word accuracies of less than 80%, as e.g. on broadcast news transcription tasks [1].

Speech recognizers may insert new words into the original sequence of spoken words and may substitute or delete others that might be essential in order to filter out the relevant portion of a document collection. Unlike text retrieval, SDR thus requires retrieval metrics that are robust towards recognition errors. In the recent past, several research groups investigated retrieval metrics that are suitable for SDR tasks [2], [3]. Surprisingly, the development of robust metrics turned out to be less difficult than expected at the beginning of the research in this field, for recognition errors seem to hardly affect retrieval performance, and this result also holds for tasks, where automatically generated transcriptions achieve word error rates of up to 40% (cf. the experimental results in Section III-A). Although, this was the unanimous result of past TREC evaluations [2], [3], the reasons are only insufficiently examined. In this article, we will conduct a probabilistic analysis of errors in SDR. To this purpose, we will propose two new error criteria that are more suitable in order to quantify the appropriateness

Manuscript received 8 April 2002 amd in revised form 30 October 2002. The authors are with the Lehrstuhl für Informatik VI, Computer Science Department, RWTH Aachen – University of Technology, D-52056 Aachen, Germany. E-mail: {w.macherey,viechtbauer,ney}@informatik.rwth-aachen.de , Fax: +49-241-8022219.

of automatically generated transcriptions for retrieval applications. The second part of this article attends to probabilistic retrieval metrics for SDR. Although, probabilistic retrieval metrics are usually better motivated in terms of a mathematically well-founded theory than their heuristic counterparts, they often suffer from lower performances. In order to compensate for this shortcoming, we propose a new statistical approach to information retrieval based on a measure for *document similarities*. Experimental results for both the error analysis and the new statistical approach are presented on the TREC-7 and TREC-8 SDR task.

The structure of this article is as follows: in Section II we start with a brief introduction to heuristic retrieval metrics. In order to improve the baseline performance, we propose a new method for query expansion. Section III is about the effect of recognition errors on retrieval performance. It includes a detailed error analysis and presents the datasets used for the experiments. In Section IV, we propose the new statistical approach to information retrieval and give detailed results of the experiments conducted. We conclude the paper with a summary in Section V.

II. HEURISTIC RETRIEVAL METRICS IN SDR

Among the proposed heuristic approaches to information retrieval the *term-frequency/inverse-document-frequency* (tf-idf) metric belongs to the best investigated retrieval metrics. Due to its simple structure in combination with a fairly well initial performance, tf-idf forms the basis for several advanced retrieval metrics. In the following section, we will give a brief introduction to tf-idf in order to introduce the terminology used in this paper and to form the basis for all further considerations.

A. Baseline methods

Let $\mathcal{D} := \{\mathbf{d}_1, \ldots, \mathbf{d}_K\}$ be a set of K documents and let $\mathbf{w} = w_1, \ldots, w_s$ denote a request, given as a sequence of s words. A retrieval system transforms $\mathbf{w}$ into a set of query terms $q_1, \ldots, q_m$ ($m \leq s$) which are used to retrieve those documents that preferably should meet the user's information need. To this purpose, all words that are of "low semantic worth" for the actual retrieval process are eliminated (*stopping*) while the residual words are reduced to their morphological stem (*stemming*), using e.g. Porter's stemming algorithm [4]. Documents are preprocessed in the same manner as the queries are. The remaining words, also referred to as *index terms*, constitute the features that describe a document or query. In the following, index terms will be denoted by d or q, if they are associated with a certain document $\mathbf{d}$ or query $\mathbf{q}$; otherwise we will use the symbol t. Let $\mathcal{T} := \{t_1, \ldots, t_T\}$ be a set of index terms and let $\mathcal{Q} := \{\mathbf{q}_1, \ldots, \mathbf{q}_L\}$ denote a set of queries. Then both documents and queries are given as sequences of index terms:

$$\begin{aligned} \mathbf{d}_k &= d_{k,1}, \ldots, d_{k,I_k}, \quad \mathbf{d}_k \in \mathcal{D} \text{ with } d_{k,i} \in \mathcal{T} \quad (1 \leq i \leq I_k) \\ \mathbf{q}_l &= q_{l,1}, \ldots, q_{l,J_l}, \quad \mathbf{q}_l \in \mathcal{Q} \text{ with } q_{l,j} \in \mathcal{T} \quad (1 \leq j \leq J_l) \end{aligned} \tag{1}$$

Each query $\mathbf{q} \in \mathcal{Q}$ partitions the document set $\mathcal{D}$ into a subset $\mathcal{D}^{\text{rel}}(\mathbf{q})$ containing all documents that are relevant w.r.t. $\mathbf{q}$, and the complementary set $\mathcal{D}^{\text{irr}}(\mathbf{q})$ containing the residual, i.e. all irrelevant documents. The number of occurrences of an index term t in a document $\mathbf{d}_k$ and a query $\mathbf{q}_l$ resp. is denoted by

$$n(t, \mathbf{d}_k) := \sum_{i=1}^{I_k} \delta(t, d_{k,i}), \qquad n(t, \mathbf{q}_l) := \sum_{j=1}^{J_l} \delta(t, q_{l,j}) \tag{2}$$

with $\delta(\cdot,\cdot)$ as the *Kronecker* function. The counts $n(t, \mathbf{d}_k)$ in Eq. (2) are also referred to as *term frequencies* of document $\mathbf{d}_k$. Using $n(t, \mathbf{d}_k)$ from Eq. (2) we define the *document frequency* $n(t)$ as the number of documents containing the index term t:

$$n(t) := \sum_{\substack{k=1 \\ n(t,\mathbf{d}_k)>0}}^{K} 1 \tag{3}$$

With the definition of the *inverse document frequency*

$$\text{idf}(t) := \log \frac{1+K}{1+n(t)} \tag{4}$$

a document specific weight $\omega(t, \mathbf{d})$ and a query specific weight $\omega(t, \mathbf{q})$ is assigned to each index term t. These weights are defined as the product over the term frequencies $n(t, \mathbf{d})$ and $n(t, \mathbf{q})$ resp. and the inverse document frequencies:

$$\omega(t, \mathbf{d}) := n(t, \mathbf{d}) \cdot \text{idf}(t), \qquad \omega(t, \mathbf{q}) := n(t, \mathbf{q}) \cdot \text{idf}(t) \tag{5}$$

Given a query $\mathbf{q}$, a retrieval system rates each document in the database whether it may meet the request or not. The result is a *ranking list* including all documents that are supposed to be relevant w.r.t. $\mathbf{q}$. To this purpose, we define a *retrieval function* f that in case of using the tf-idf metric is defined as the product over all weights of index terms occurring in $\mathbf{q}$ as well as in $\mathbf{d}$, normalized by the length of the query $\mathbf{q}$ and the document $\mathbf{d}$:

$$f(\mathbf{q}, \mathbf{d}) := \frac{\sum_{t \in \mathcal{T}} \omega(t, \mathbf{q}) \cdot \omega(t, \mathbf{d})}{\sqrt{\sum_{t \in \mathcal{T}} n^2(t, \mathbf{q})} \cdot \sqrt{\sum_{t \in \mathcal{T}} n^2(t, \mathbf{d})}} \tag{6}$$

The value of $f(\mathbf{q}, \mathbf{d})$ is called *retrieval status value* (RSV). The evaluation of $f(\mathbf{q}, \mathbf{d})$ for all documents $\mathbf{d} \in \mathcal{D}$ induces a ranking according to which the documents are compiled to a list that is sorted in descending order. The higher the RSV of a document, the better it may meet the query and the more important it may be for the user.

B. Advanced retrieval metrics

Based on the tf-idf metric, several modifications were proposed in literature, leading e.g. to the OKAPI metrics [5] as well as the SMART-1 and the SMART-2 metric [6]. The baseline results conducted for this paper use the following version of the SMART-2 metric. Here, the inverse document frequencies are given by:

$$\mathrm{idf}(t) := \log\left\lfloor \frac{K}{n(t)} \right\rfloor \tag{7}$$

Note that due to the floor operation in Eq. (7) a term weight will be zero if it occurs in more than half of the documents. According to [7], each index term t in a document $\mathbf{d}$ is associated with a weight $g(t,\mathbf{d})$ that depends on the ratio of the logarithm of the term frequency $n(t,\mathbf{d})$ to the logarithm of the average term frequency $\overline{n}(\mathbf{d})$

$$g(t,\mathbf{d}) := \begin{cases} \big[1+\log n(t,\mathbf{d})\big] \big/ \big[1+\log \overline{n}(\mathbf{d})\big] & \text{if } t \in \mathbf{d} \\ 0 & \text{if } t \notin \mathbf{d} \end{cases} \tag{8}$$

with $\log 0 := 0$ and

$$\overline{n}(\mathbf{d}) = \frac{\sum_{t\in\mathcal{T}} n(t,\mathbf{d})}{\sum_{t\in\mathcal{T}:n(t,\mathbf{d})>0} 1} \tag{9}$$

The logarithms in Eq. (8) prevent documents with high term frequencies from dominating those with low term frequencies. In order to obtain the final term weights, $g(t,\mathbf{d})$ is divided by a linear combination between a pivot element c and the number of singletons $n_1(\mathbf{d})$ in document $\mathbf{d}$:

$$\omega(t,\mathbf{d}) := \frac{g(t,\mathbf{d})}{(1-\lambda)\cdot c + \lambda \cdot n_1(\mathbf{d})} \tag{10}$$

with $\lambda = 0.2$ and

$$c = \frac{1}{K}\sum_{k=1}^{K} n_1(\mathbf{d}_k) \quad \text{and} \quad n_1(\mathbf{d}) := \sum_{t\in\mathcal{T}:n(t,\mathbf{d})=1} 1 \tag{11}$$

Unlike tf-idf, only query terms are weighted with the inverse document frequency $\mathrm{idf}(t)$:

$$\omega(t,\mathbf{q}) = \big[1+\log n(t,\mathbf{q})\big]\cdot \mathrm{idf}(t) \tag{12}$$

Now, we can define the SMART-2 retrieval function as the product over the document and query specific index term weights:

$$f(\mathbf{q},\mathbf{d}) = \sum_{t\in\mathcal{T}} \omega(t,\mathbf{q})\cdot\omega(t,\mathbf{d}) \tag{13}$$

C. Improving retrieval performance

Often, the retrieval effectiveness can be improved using interactive search techniques such as *relevance feedback* methods. Retrieval systems providing relevance feedback conduct a preliminary search and present the top ranked documents to the user who has to rate each document whether it meets his information need or not. Based on this relevance judgment, the original query vector is modified in the following way. Let $\widetilde{\mathcal{D}}^{\mathrm{rel}}(\mathbf{q})$ be the subset of top ranked documents rated as relevant and let $\widetilde{\mathcal{D}}^{\mathrm{irr}}(\mathbf{q})$ denote the subset of irrelevant retrieved documents. Further let $\mathbf{e_d}$ denote the document $\mathbf{d}$ embedded into a T-dimensional vector $\mathbf{e_d} = \big(n(t_1,\mathbf{d}),\ldots,n(t_T,\mathbf{d})\big)^\top$ and let $\mathbf{e_q} = \big(n(t_1,\mathbf{q}),\ldots,n(t_T,\mathbf{q})\big)^\top$ denote the vector embedding of the query $\mathbf{q}$. Then, the difference vector $\boldsymbol{\rho_q}$ defined by

$$\boldsymbol{\rho_q} = \frac{1}{|\widetilde{\mathcal{D}}^{\mathrm{rel}}(\mathbf{q})|}\cdot \sum_{\mathbf{d}\in\widetilde{\mathcal{D}}^{\mathrm{rel}}(\mathbf{q})} \mathbf{e_d} \;-\; \frac{1}{|\widetilde{\mathcal{D}}^{\mathrm{irr}}(\mathbf{q})|}\cdot \sum_{\mathbf{d}\in\widetilde{\mathcal{D}}^{\mathrm{irr}}(\mathbf{q})} \mathbf{e_d} \tag{14}$$

connects the centroids of both document subsets. Therefore, it can be used in order to shift the original query vector $\mathbf{e_q}$ towards the cluster of relevant documents, resulting in a new query vector $\widetilde{\mathbf{e}}_\mathbf{q}$ (cf. Fig. 1):

$$\widetilde{\mathbf{e}}_\mathbf{q} = (1-\gamma)\cdot \mathbf{e_q} \;+\; \gamma\cdot \boldsymbol{\rho}_\mathbf{q} \qquad (0 \le \lambda \le 1) \tag{15}$$

This method is also known as *query expansion* and the *Rocchio algorithm* [8] counts among the best known implementations of this idea, although there are many others as well [9], [10], [11]. Assuming that the r top ranked documents of the preliminary search are (most likely) relevant, interactive search techniques can be automated by setting $\widetilde{\mathcal{D}}^{\mathrm{rel}}(\mathbf{q})$ to the first r retrieved documents, whereas $\widetilde{\mathcal{D}}^{\mathrm{irr}}(\mathbf{q})$ is set to $\emptyset$. However, since the effectiveness of a preliminary retrieval process may decrease due to recognition errors, query expansion is often performed on secondary document collections, e.g. news paper articles that are kept apart from the actual retrieval corpus. This technique is very effective, but at the same time it requires significantly more resources due to the additional indexing and storage costs of the supplementary database. Therefore, we focus on a new method for query expansion that solely uses the actual retrieval corpus while preserving robustness towards recognition errors. The approach comprises the following three steps:

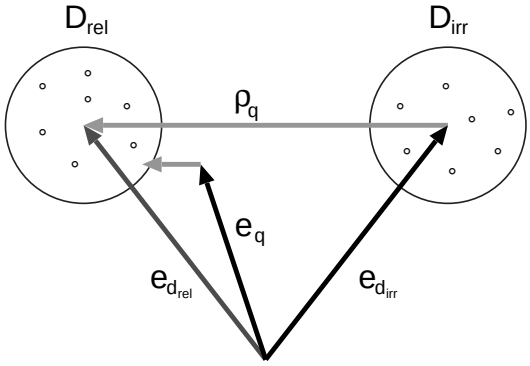

Fig. 1. Principle of query expansion: using the difference vector $\boldsymbol{\rho_q}$, the original query vector $\mathbf{e_q}$ is shifted towards the subset of relevant documents.

1. Perform a preliminary retrieval using SMART-2 with $\pi : \{1,\ldots,K\} \rightarrow \{1,\ldots,K\}$ induced by the ranking list so that $f(\mathbf{q},\mathbf{d}_{\pi(1)}) \geq \ldots \geq f(\mathbf{q},\mathbf{d}_{\pi(K)})$ holds.
2. Determine the query expansion vector $\widehat{\mathbf{e}}_{\mathbf{q}}$ defined as the sum over the expansion vectors $\mathbf{v}_{\mathbf{q}}(\mathbf{d})$ of the r top ranked documents $\mathbf{d}_{\pi(1)},\ldots,\mathbf{d}_{\pi(r)}$ $(r \leq K)$

$$\widehat{\mathbf{e}}_{\mathbf{q}} := \sum_{\mathbf{d}\in\mathcal{D}\,:\, f(\mathbf{q},\mathbf{d}_{\pi(1)})\geq f(\mathbf{q},\mathbf{d})\geq f(\mathbf{q},\mathbf{d}_{\pi(r)})} \frac{\mathbf{v}_{\mathbf{q}}(\mathbf{d})}{\sqrt{||\mathbf{v}_{\mathbf{q}}(\mathbf{d})||^2}} \quad (16)$$

with the ith component $(1 \leq i \leq T)$ of $\mathbf{v}_{\mathbf{q}}(\mathbf{d})$ given by

$$v^i_{\mathbf{q}}(\mathbf{d}) := \begin{cases} g(t_i,\mathbf{d}) \cdot \mathrm{idf}(t_i) \cdot \log n(t_i,\mathbf{d}) & \text{if } t_i \notin \mathbf{q} \\ 0 & \text{if } t_i \in \mathbf{q} \end{cases} \quad (17)$$

3. The new query vector $\widetilde{\mathbf{e}}_{\mathbf{q}}$ is defined by

$$\widetilde{\mathbf{e}}_{\mathbf{q}} = \mathbf{e}_{\mathbf{q}} + \gamma \cdot \sqrt{||\mathbf{e}_{\mathbf{q}}||^2} \cdot \frac{\widehat{\mathbf{e}}_{\mathbf{q}}}{\sqrt{||\widehat{\mathbf{e}}_{\mathbf{q}}||^2}} \quad (18)$$

III. Analysis of Recognition Errors and Retrieval Performance

Switching from manual to recognized transcriptions raises the question of robustness of retrieval metrics towards recognition errors. Automatic speech recognition (ASR) systems may insert new words into the original sequence of spoken words while substituting or deleting others that might be essential in order to filter out the relevant portion of a document collection. In ASR, the performance is usually measured in terms of *word error rate* (WER). The WER is defined as the *Levenshtein* or edit distance, which is the minimal number of insertions (ins), deletions (del) and substitutions (sub) of words necessary to transform the spoken sentence into the recognized sentence. The relative WER is defined by:

$$\mathrm{WER} := \sum_{k=1}^{K} \frac{\mathrm{sub}_k + \mathrm{ins}_k + \mathrm{del}_k}{N} \quad (19)$$

Here, N is the total number of words in the reference transcriptions of the document collection $\mathcal{D}$. The computation of the WER requires an alignment of the spoken sentence with the recognized sentence. Thus, the order of words is explicitly taken into account.

A. Tasks and experimental results

Experiments for the investigation on the effect of recognition errors on retrieval performance were carried out on the TREC-7 and the TREC-8 SDR task using manually segmented stories [3]. The TREC-7 task comprises 2866 documents and 23 test queries. The TREC-8 task comprises 21745 spoken documents and 50 test queries. Table I summarizes some corpus statistics.

TABLE I
CORPUS STATISTICS FOR THE TREC-7 AND THE TREC-8 SPOKEN DOCUMENT RETRIEVAL TASK.

	TREC-7			TREC-8		
	all	rel.	irr.	all	rel.	irr.
# documents	2866	348	2518	21745	1679	20066
# queries	23	—	—	50	—	—
avg. doc. length	267.4	580.1	265.5	169.6	283.9	169.4

Recognition results on the TREC-7 SDR tasks were produced using the RWTH large vocabulary continuous speech recognizer (LVCSR) [12]. The recognizer uses a time-synchronous beam search algorithm based on the concept of word-dependent tree copies and integrates the trigram language-model constraints in a single pass. Besides acoustic and histogram pruning a look-ahead technique of the language model probabilities is utilized [13]. Recognition results were produced using gender independent models. Neither speaker-adaptive nor any normalization methods were applied. Every nine consecutive feature vectors, each consisting of 16 cepstral coefficients, are spliced and mapped onto a 45 dimensional feature vector using a *linear discriminant analysis* (LDA). The segmentation of the audio stream into speech and non-speech segments is based on a Gaussian mixture distribution model.

Table II shows the effect of recognition errors on retrieval performance, measured in terms of *mean average precision* (MAP) [14] for different retrieval metrics on the TREC-7 SDR task. Even though, the WER of the recognized transcriptions is 32.5%, the retrieval performance decreases by only 9.9% relative using the SMART-2 metric in comparison with the original, i.e. the manually generated transcriptions. The relative loss is even smaller (approx. 5% relative) if the new query expansion method is used.

Similar results could be observed on the TREC-8 corpus. Unlike the experiments conducted on the TREC-7 SDR task, we made use of the recognition outputs of the

TABLE II
RETRIEVAL EFFECTIVENESS MEASURED IN TERMS OF MAP ON THE TREC-7 AND THE TREC-8 SDR TASK. ALL WERS WERE DETERMINED WITHOUT NIST RE-SCORING. THE NUMBERS IN PARENTHESES INDICATE THE RELATIVE CHANGE BETWEEN TEXT AND SPEECH BASED RESULTS.

metric		MAP[%] TREC-7	MAP[%] TREC-8	
text	tf-idf	42.1	47.6	
	SMART-2	46.6	49.6	
	q-expansion	53.4	57.5	
speech	tf-idf	35.3 (-16.2%)	41.3 (-13.2%)	42.0 (-11.8%)
	SMART-2	42.0 (-9.9%)	43.1 (-13.1%)	42.1 (-15.1%)
	q-expansion	50.7 (-5.1%)	50.0 (-13.0%)	49.8 (-13.4%)
	WER[%]	32.5 (RWTH)	38.4 (Byblos)	40.3 (Dragon)

Byblos "Rough 'N Ready" LVCSR system [15] and the Dragon LVCSR system [16]. Here, the retrieval performance decreases by only 13.1% relative using the SMART-2 metric in combination with the recognition outputs of the Byblos speech recognizer and by 15.1% relative using the Dragon speech recognition outputs. Note that in both cases the WER is approximately 40%, i.e. almost every second word was misrecognized. Using the new query expansion method, the relative performance loss is nearly constant, i.e. the transcriptions as produced by the Byblos speech recognizer cause a performance loss of 13.0% relative whereas the transcriptions generated by the Dragon system cause a degradation of 13.4% relative.

B. Alternative error measures

Since most retrieval metrics usually disregard word orders, the WER is certainly not suitable in order to quantify the quality of recognized transcriptions for retrieval applications. A more reasonable error measure is given by the *term error rate* (TER) as proposed in [17]:

$$\mathrm{TER} := \frac{1}{K} \cdot \sum_{k=1}^{K} \frac{\sum_{t \in \mathcal{T}} \left| n(t, \widehat{\mathbf{d}}_k) - n(t, \mathbf{d}_k) \right|}{I_k} \tag{20}$$

As before, I_k denotes the number of index terms in the reference document $\mathbf{d}_k$ and $n(t, \mathbf{d}_k)$ is the original term frequency. $n(t, \widehat{\mathbf{d}}_k)$ denotes the term frequency of the term t in the recognized transcription $\widehat{\mathbf{d}}_k$. Note that a substitution error according to the WER produces two errors in terms of the TER, since it not only misses a correct word, but also introduces a spurious one. Consequently, we have to count substitutions twice in order to compare both error measures. Nevertheless, the alignment the WER computation is based on, must still be determined using uniform costs, i.e. substitutions are counted once. Using the definitions

$$\mathrm{del}_t(\mathbf{d}, \widehat{\mathbf{d}}) := \begin{cases} n(t, \mathbf{d}) - n(t, \widehat{\mathbf{d}}) & : \quad n(t, \widehat{\mathbf{d}}) < n(t, \mathbf{d}) \\ 0 & : \quad \text{otherwise} \end{cases}$$

$$\mathrm{ins}_t(\mathbf{d}, \widehat{\mathbf{d}}) := \begin{cases} n(t, \widehat{\mathbf{d}}) - n(t, \mathbf{d}) & : \quad n(t, \widehat{\mathbf{d}}) > n(t, \mathbf{d}) \\ 0 & : \quad \text{otherwise} \end{cases}$$

the TER can be rewritten as:

$$\mathrm{TER} = \frac{1}{K} \sum_{k=1}^{K} \sum_{t \in \mathcal{T}} \frac{\mathrm{del}_t(\mathbf{d}_k, \widehat{\mathbf{d}}_k) \; + \; \mathrm{ins}_t(\mathbf{d}_k, \widehat{\mathbf{d}}_k)}{I_k} \tag{21}$$

Since the contributions of term frequencies to term weights are often diminished by the application of logarithms (cf. Eq. (8)), the number of occurrences of an index term within a document $\mathbf{d}$ is of less importance than the fact whether a term *does* occur in $\mathbf{d}$ or not. Therefore, we propose the *indicator error rate* (IER) that is defined by:

$$\mathrm{IER} := \frac{1}{K} \cdot \sum_{k=1}^{K} \frac{\left| \mathcal{T}_{\mathbf{d}_k} \setminus \mathcal{T}_{\widehat{\mathbf{d}}_k} \right| + \left| \mathcal{T}_{\widehat{\mathbf{d}}_k} \setminus \mathcal{T}_{\mathbf{d}_k} \right|}{\left| \mathcal{T}_{\mathbf{d}_k} \right|} \tag{22}$$

with

$$\mathcal{T}_{\mathbf{d}_k} := \left\{ d_{k,1}, \cdots, d_{k,I_k} \right\} \qquad (1 \le k \le K) \tag{23}$$

The IER discards term frequencies and measures the number of index terms that were missed or wrongly added during recognition. If we transfer the concepts *recall* and *precision* to pairs of documents we will obtain a motivation for the IER. To this purpose, we define

$$\mathrm{recall}(\mathbf{d}, \widehat{\mathbf{d}}) := \frac{\left| \mathcal{T}_{\mathbf{d}} \cap \mathcal{T}_{\widehat{\mathbf{d}}} \right|}{\left| \mathcal{T}_{\mathbf{d}} \right|}, \qquad \mathrm{prec}(\mathbf{d}, \widehat{\mathbf{d}}) := \frac{\left| \mathcal{T}_{\mathbf{d}} \cap \mathcal{T}_{\widehat{\mathbf{d}}} \right|}{\left| \mathcal{T}_{\widehat{\mathbf{d}}} \right|}$$

Note that a high recall means that the recognized transcription $\widehat{\mathbf{d}}$ contains many index terms of the reference transcription $\mathbf{d}$. A low precision means that the recognized transcription contains many index terms that do not occur in the reference transcription. Both the recall and precision error are given by:

$$1 - \mathrm{recall}(\mathbf{d}, \widehat{\mathbf{d}}) = \frac{\left| \mathcal{T}_{\mathbf{d}} \setminus \mathcal{T}_{\widehat{\mathbf{d}}} \right|}{\left| \mathcal{T}_{\mathbf{d}} \right|}, \quad 1 - \mathrm{prec}(\mathbf{d}, \widehat{\mathbf{d}}) = \frac{\left| \mathcal{T}_{\widehat{\mathbf{d}}} \setminus \mathcal{T}_{\mathbf{d}} \right|}{\left| \mathcal{T}_{\widehat{\mathbf{d}}} \right|}$$

If we assume both the reference and the recognized documents to be of the same size, i.e. $|\mathcal{T}_{\mathbf{d}}| \approx |\mathcal{T}_{\widehat{\mathbf{d}}}|$ which can be justified by the fact that language model scaling factors are usually set to values ensuring balanced numbers of deletions and insertions, we obtain the following interpretation of the IER:

$$\begin{aligned} \mathrm{IER} &= \frac{1}{K} \cdot \sum_{k=1}^{K} \frac{\left| \mathcal{T}_{\mathbf{d}_k} \setminus \mathcal{T}_{\widehat{\mathbf{d}}_k} \right| + \left| \mathcal{T}_{\widehat{\mathbf{d}}_k} \setminus \mathcal{T}_{\mathbf{d}_k} \right|}{|\mathcal{T}_{\mathbf{d}_k}|} \\ &\approx \frac{1}{K} \cdot \sum_{k=1}^{K} \left[1 - \frac{\left| \mathcal{T}_{\mathbf{d}_k} \setminus \mathcal{T}_{\widehat{\mathbf{d}}_k} \right|}{\left| \mathcal{T}_{\mathbf{d}_k} \right|} \; + \; 1 - \frac{\left| \mathcal{T}_{\widehat{\mathbf{d}}_k} \setminus \mathcal{T}_{\mathbf{d}_k} \right|}{\left| \mathcal{T}_{\widehat{\mathbf{d}}_k} \right|} \right] \\ &= \frac{1}{K} \cdot \sum_{k=1}^{K} \left[2 - \mathrm{recall}(\mathbf{d}_k, \widehat{\mathbf{d}}_k) - \mathrm{prec}(\mathbf{d}_k, \widehat{\mathbf{d}}_k) \right] \end{aligned}$$

Table III shows the error rates obtained on the TREC-7 SDR task for the three error measures WER, TER, and IER. Note that substitution errors are counted twice in order to be comparable with the TER. The initial WER thus obtained is 52.8% on the whole document collection, whereas TER leads to an initial error rate of 44.6%. So far, we have not yet taken into account the effect of document preprocessing steps, i.e. stopping and stemming. If we consider index terms only, TER decreases to 42.8%. Moreover, we can restrict the index terms to query terms only. Thus, TER decreases to 29.5%. Note that this magnitude will correspond to a WER of 17.4% if we convert TER into WER using the initial ratio of deletions, insertions, and substitutions of 4.8 : 4.7 : 21.6. Finally, we can apply the indicator error measure which leads to an IER of 19.5%, thus corresponding to a WER of 17.4%. Similar

TABLE III

WERs, TERs, IERs measured with the RWTH speech recognizer on the Trec-7 corpus for varying preprocessing stages. Note that the substitutions are counted twice for the accumulated error rates of the WER criterion.

		Trec-7			Trec-7 + stop + stem			+ stop + stem, queries only		
	documents	all	relevant	irrelevant	all	relevant	irrelevant	all	relevant	irrelevant
WER[%]	deletions	4.8	3.9	4.9	8.5	6.3	8.8	11.1	8.2	11.5
	insertions	4.7	4.1	4.8	2.6	2.4	2.6	8.7	6.7	9.0
	substitutions	21.6	18.4	22.1	17.0	14.2	17.3	5.3	4.7	5.4
	error rate*	**52.8**	44.7	53.9	45.0	37.2	46.0	30.3	24.4	31.2
TER[%]	deletions	21.8	17.4	22.4	24.0	19.2	24.6	12.0	10.8	12.2
	insertions	22.8	17.9	23.5	18.9	15.5	19.3	17.5	10.8	18.4
	error rate	**44.6**	35.3	45.9	**42.8**	34.7	43.9	**29.5**	21.5	30.6
IER[%]	deletions	16.3	13.9	16.6	17.4	14.2	17.9	8.8	7.0	9.0
	insertions	16.3	14.2	16.5	15.1	13.6	15.3	10.7	8.4	11.0
	error rate	32.5	28.1	33.1	32.5	27.8	33.2	**19.5**	**15.5**	20.0

TABLE IV

Summary of different error measures on the Trec-7 and Trec-8 SDR task. Substitution errors (sub) are counted once (sub 1×) or twice (sub 2×), respectively.

doc.	error measure	Trec-7 RWTH	Trec-8 Byblos	Trec-8 Dragon
all	WER[%] (sub 1×)	32.5	38.4	40.3
	(sub 2×)	52.8	60.3	61.3
	TER[%]	44.6	52.2	53.2
	+stop +stem	42.8	48.8	49.2
	q-terms only	29.5	34.8	36.7
	IER[%] q-terms only	19.5	22.3	23.4
rel.	IER[%] q-terms only	15.5	18.0	18.7

results were observed on the Trec-8 SDR task, using the recognition outputs of the Byblos and the Dragon speech recognition system, respectively (cf. Tables VIII and IX). Table IV summarizes the most important error rates of the Tables III, VIII, and IX.

For each error measure we can determine the accuracy rate which is given by $\max(1 - \mathrm{ER},\ 0)$ where ER is the WER, the TER, or the IER, respectively. Assuming a linear dependency of the retrieval effectiveness on the accuracy rate, we can compute the squared empirical correlation between the MAP obtained on the recognized documents and the product over the accuracy rate and the MAP obtained on the reference documents. Table V shows the correlation coefficients thus computed. The computation of the accuracy rates refer to the ninth column of the Tables III, VIII and IX, i.e. all documents were stopped and stemmed beforehand and reduced to query terms. Substitutions were counted only once in order to determine the word accuracies. Among the proposed error measures, the IER seems to best correlate with the retrieval effectiveness. However, the amount of data is still too small and further experiments will be necessary to prove this proposition.

TABLE V

Squared empirical correlation between the MAP obtained on the recognized documents and the MAP obtained on the reference documents multiplied with the word accuracy rate (WA), the term accuracy rate (TA) and the indicator accuracy rate (IA), respectively.

accuracy rate	tf-idf	Smart-2	q-expansion
WA	0.741	0.323	0.010
TA	0.475	0.007	0.567
IA	0.937	0.845	0.688

C. Further discussion

In this section we will investigate the magnitude of the performance loss from a theoretical point of view. To this purpose, we consider the retrieval process in detail. When a user addresses a query to a retrieval system, each document in the database is rated according to its RSV. The induced ranking list determines a permutation π of the documents that can be mapped onto a vector indicating whether the document $\mathbf{d}_i$ at position $\pi(i)$ is relevant w.r.t. $\mathbf{q}$ or not. Let f be a retrieval function. Then, the application of f to a document collection $\mathcal{D}$ given a query $\mathbf{q}$ leads to the permutation $f_{\mathbf{q}}(\mathcal{D}) = \left(\mathbf{d}_{\pi(1)}, \mathbf{d}_{\pi(2)}, \ldots, \mathbf{d}_{\pi(K)}\right)$ with π induced by the following order:

$$f(\mathbf{q}, \mathbf{d}_{\pi(1)}) \geq f(\mathbf{q}, \mathbf{d}_{\pi(2)}) \geq \ldots \geq f(\mathbf{q}, \mathbf{d}_{\pi(K)})$$

With the definition of the indicator function

$$\mathcal{I}_{\mathbf{q}}(\mathbf{d}) := \begin{cases} 1 & \text{if } \mathbf{d} \text{ is relevant w.r.t. } \mathbf{q} \\ 0 & \text{otherwise} \end{cases}$$

the ranking list can be mapped onto a binary vector:

$$\mathcal{I}_{\mathbf{q}} \begin{pmatrix} \mathbf{d}_{\pi(1)} \\ \mathbf{d}_{\pi(2)} \\ \mathbf{d}_{\pi(3)} \\ \vdots \\ \mathbf{d}_{\pi(n)} \\ \mathbf{d}_{\pi(n+1)} \\ \vdots \\ \mathbf{d}_{\pi(K)} \end{pmatrix} \mapsto \begin{pmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Even though the deterioration of transcriptions as caused by recognition errors may change the indicator vector, a performance loss will only occur if the RSVs of relevant documents fall below the RSVs of irrelevant documents. Note that among the four possible cases of local exchange operations between documents, i.e. $\mathcal{I}_{\mathbf{q}}(\mathbf{d}_{\pi(i)}) \in \{0,1\}$ changes its position with $\mathcal{I}_{\mathbf{q}}(\mathbf{d}_{\pi(j)}) \in \{0,1\}$ $(i \neq j)$, only one case can cause a performance loss. Interestingly, it is possible to specify an upper bound for the probability that two documents $\mathbf{d}_i$ and $\mathbf{d}_j$ with $f(\mathbf{q},\mathbf{d}_i) > f(\mathbf{q},\mathbf{d}_j)$ will change their relative order if they are deteriorated by recognition errors, i.e. $f(\mathbf{q},\widehat{\mathbf{d}}_i) < f(\mathbf{q},\widehat{\mathbf{d}}_j)$ shall hold for the recognized documents $\widehat{\mathbf{d}}_i$ and $\widehat{\mathbf{d}}_j$. According to [18], this upper bound is given by:

$$P\big(f(\mathbf{q},\widehat{\mathbf{d}}_i) > f(\mathbf{q},\widehat{\mathbf{d}}_j) \,\big|\, f(\mathbf{q},\mathbf{d}_i) < f(\mathbf{q},\mathbf{d}_j)\big) \leq \sum_{t\in\mathcal{T}} \frac{n^2(t,\mathbf{q}) \cdot \Big[\sigma\big(n(t,\mathbf{d}_i)\big)/I_i + \sigma\big(n(t,\mathbf{d}_j)\big)/I_j\Big]}{\Delta_{i,j}^2(\mathbf{q})}$$

with

$$\Delta_{i,j}(\mathbf{q}) := E\big[f(\mathbf{q},\widehat{\mathbf{d}}_i)\big] - E\big[f(\mathbf{q},\widehat{\mathbf{d}}_j)\big]$$

$$E\big[f(\mathbf{q},\widehat{\mathbf{d}})\big] := \sum_{t\in\mathbf{q}} \frac{n(t,\mathbf{q}) \cdot \big[p_c(t) - p_e(t)\big] \cdot n(t,\mathbf{d})}{l(\mathbf{d})} + p_e(t)$$

$$\sigma\big[n(t,\widehat{\mathbf{d}})\big] := \Big[p_c(t)\cdot\big[1-p_c(t)\big] - p_e(t)\cdot\big[1-p_e(t)\big]\Big] \cdot n(t,\mathbf{d}) + l(\mathbf{d})\cdot p_e(t)$$

Here, $p_c(t)$ denotes the probability that t is correctly recognized and $p_e(t)$ is the probability that t is recognized even though τ $(\tau \neq t)$ was spoken. $l(\mathbf{d})$ is a document specific length normalization that depends on the used retrieval metric. Thus, the upper bound for the probability of changing the order of two documents is vanishing for increasing document lengths [14, p. 135]. In particular this means that the relevant documents of the TREC-7 and the TREC-8 corpus are less affected by recognition errors than irrelevant documents since the average length of relevant documents is substantially larger than the average length of irrelevant documents (cf. Table I).

Now, let $\pi_0 : \{1,\ldots,K\} \rightarrow \{1,\ldots,K\}$ denote a permutation of the documents so that $f(\mathbf{q},\mathbf{d}_{\pi_0(1)}) > \ldots > f(\mathbf{q},\mathbf{d}_{\pi_0(K)})$ holds for a query $\mathbf{q}$. Then, we can define a matrix $\mathbf{A} \in I\!R_+^{K\times K}$ with elements

$$a_{ij} := P\big(f(\mathbf{q},\widehat{\mathbf{d}}_{\pi_0(i)}) < f(\mathbf{q},\widehat{\mathbf{d}}_{\pi_0(j)}) \,\big|\, f(\mathbf{q},\mathbf{d}_{\pi_0(i)}) > f(\mathbf{q},\mathbf{d}_{\pi_0(j)})\big)$$

At the beginning, $\mathbf{A}$ is an upper triangular matrix whose diagonal elements are zero. Since exchanges between relevant documents and exchanges between irrelevant documents do not affect the retrieval performance, each matrix element a_{ij} will be set to 0 if $\{\mathbf{d}_{\pi_0(i)},\mathbf{d}_{\pi_0(j)}\} \subseteq \mathcal{D}^{\mathrm{rel}}(\mathbf{q})$ or $\{\mathbf{d}_{\pi_0(i)},\mathbf{d}_{\pi_0(j)}\} \subseteq \mathcal{D}^{\mathrm{irr}}(\mathbf{q})$. Then, the expectation of the ranking, i.e. the permutation π maximizing the MAP of the recognized documents can be determined according to the following algorithm using a greedy policy:

π:=π_0;
for i:=1 to K do begin
  π_i(i):=argmax$_j\{a_{ij}\}$;
  for k:=1 to K do begin if(k≠i) π_i(k):=k; end;
  $a_{i,\pi_i(i)}$:=0;
  π:=$\pi_i \circ \pi$;
end;

The sequence of permutations $\pi_K \circ \ldots \circ \pi_1 \circ \pi_0$ defines a sequence of re-orderings that corresponds with the expectation of the new ranking. The expectation will maximize the likelihood if the documents in the database are pairwise stochastically independent.

IV. Probabilistic Approaches to IR

Besides heuristically motivated retrieval metrics, several probabilistic approaches to information retrieval were proposed and investigated over the past years. The methods range from binary independence retrieval models [19] over language model based approaches [20] up to methods based on statistical machine translation [21]. The starting point of most probabilistic approaches to IR is the *a-posteriori* probability $p(\mathbf{d}|\mathbf{q})$ of a document $\mathbf{d}$, given a query $\mathbf{q}$. The posterior probability can directly be interpreted as RSV. In contrast to many heuristic retrieval models, RSVs of probabilistic approaches are thus always normalized and even comparable between different queries. Often, the posterior probability $p(\mathbf{d}|\mathbf{q})$ is denoted by $p(\mathbf{d}, b \in \{\mathrm{rel},\mathrm{irr}\}|\mathbf{q})$, with the random variable b indicating the relevance of $\mathbf{d}$ w.r.t. $\mathbf{q}$. However, since we consider non-interactive retrieval methods only, b is not observable and therefore obsolete, since it cannot affect the retrieval process. The a-posteriori probability can be rewritten as:

$$p(\mathbf{d}|\mathbf{q}) = \frac{p(\mathbf{d})\cdot p(\mathbf{q}|\mathbf{d})}{\sum_{\widetilde{\mathbf{d}}\in\mathcal{D}} p(\widetilde{\mathbf{d}})\cdot p(\mathbf{q}|\widetilde{\mathbf{d}})} \tag{24}$$

A document maximizing Eq. (24) is determined using Bayes' decision rule:

$$\mathbf{q} \mapsto r(\mathbf{q}) = \operatorname*{argmax}_{\mathbf{d}} \big\{ p(\mathbf{q}|\mathbf{d})\cdot p(\mathbf{d}) \big\} \tag{25}$$

This decision rule is known to be optimal with respect to the expected number of decision errors if the required distributions are known [22]. However, as neither $p(\mathbf{q}|\mathbf{d})$ nor $p(\mathbf{d})$ are known in practical situations, it is necessary to choose models for the respective distributions and estimate their parameters using suitable training data. Note that

Eq. (25) can easily be extended to a ranking by determining not only the document maximizing $p(\mathbf{d}|\mathbf{q})$, but also by compiling a list that contains all documents sorted in descending order w.r.t. their posterior probability.

In the recent past, several probabilistic approaches to information retrieval were proposed and evaluated. In [21] the authors describe a method based on statistical machine translation. A query is considered as a sequence of keywords extracted from an imaginary document that best meets the user's information need. Pairs of queries and documents are considered as bilingual annotated texts, where the objective of finding relevant documents is ascribed to a translation of a query (source language) into a document (target language). Experiments were carried out on various TREC tasks. Using the IBM-1 translation model [23] as well as a simplified version called IBM-0 the obtained retrieval effectiveness outperformed the tf-idf metric.

The approach presented in [24] makes use of multi state hidden Markov models (HMM) to interpolate document specific language models with a background language model. The background language model that is estimated on the whole document collection is used in order to smooth the probabilities of unseen index terms in the document specific language models. Experiments performed on the TREC-7 ad hoc retrieval task showed better results than tf-idf.

In [25] the authors investigate an advanced version of the Markovian approach as proposed by [24]. Experiments conducted on the TREC-7 and TREC-8 SDR tasks achieve a retrieval effectiveness that is comparable with the OKAPI metric, but does not outperform the SMART-2 results.

Even though many probabilistic retrieval metrics are able to outperform basic retrieval metrics as for example tf-idf, they usually do not achieve the effectiveness of advanced heuristic retrieval metrics such as SMART-2 or OKAPI. In particular for SDR tasks, probabilistic metrics often turned out to be less robust towards recognition errors than their heuristic counterparts. To compensate for this, we propose a new statistical approach to information retrieval that is based on document similarities [26].

A. Probabilistic retrieval using document representations

A fundamental difficulty in statistical approaches to information retrieval is the fact that typically a rare index term is well suited to filter out a document. On the other hand, a reliable estimation of distribution parameters requires that the underlying events, i.e. index terms are observed as frequently as possible. Therefore, it is necessary to properly smooth the distributions. In our case, document specific term probabilities $p(t\,|\mathbf{d})$ are smoothed with term probabilities of documents that are similar to $\mathbf{d}$. The similarity measure is based on *document representations* which in the simplest case can be document specific histograms of the index terms.

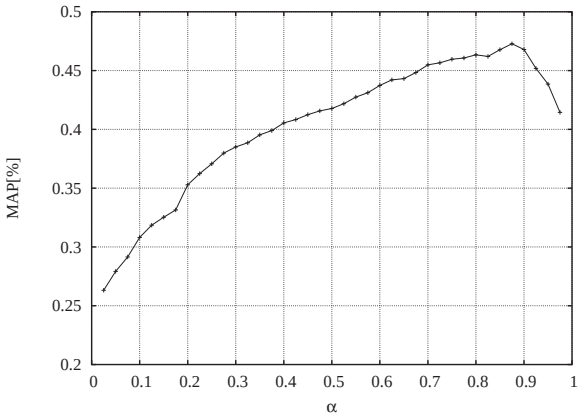

Fig. 2. MAP as a function of the interpolation parameter α with fixed $\beta = 0.300$ on the reference transcriptions of the TREC-7 SDR task.

The starting point of our approach is the joint probability $p(\mathbf{q}, \mathbf{d})$ of a query $\mathbf{q}$ and a document $\mathbf{d}$:

$$p(\mathbf{q}, \mathbf{d}) = \prod_{j=1}^{|\mathbf{q}|} p(q_j, \mathbf{d}\,|\,q_1^{j-1}) \tag{26}$$

$$= \prod_{j=1}^{|\mathbf{q}|} p(q_j, \mathbf{d}) \tag{27}$$

Here, $|\mathbf{q}|$ denotes the number of index terms in $\mathbf{q}$. The conditional probabilities $p(q_j, \mathbf{d}|\, q_1^{j-1})$ in Eq. (26) are assumed to be independent of the predecessor terms q_1^{j-1}. Document representations are now introduced via a hidden variable $\mathbf{r}$ that runs over a finite set $\mathbf{R}$ of document representations:

$$p(\mathbf{q}, \mathbf{d}) = \prod_{j=1}^{|\mathbf{q}|} \sum_{\mathbf{r}\in\mathbf{R}} p(q_j, \mathbf{d}, \mathbf{r}) \tag{28}$$

$$= \prod_{j=1}^{|\mathbf{q}|} \sum_{\mathbf{r}\in\mathbf{R}} p(q_j\,|\mathbf{r}) \cdot p(\mathbf{d}|\mathbf{r}) \cdot p(\mathbf{r}) \tag{29}$$

$$= \prod_{j=1}^{|\mathbf{q}|} \sum_{\mathbf{r}\in\mathbf{R}} p(q_j\,|\mathbf{r}) \cdot \prod_{i=1}^{|\mathbf{d}|} p(d_i\,|\mathbf{r}, d_1^{i-1}) \cdot p(\mathbf{r}) \tag{30}$$

$$= \prod_{j=1}^{|\mathbf{q}|} \sum_{\mathbf{r}\in\mathbf{R}} p(q_j\,|\mathbf{r}) \cdot \prod_{i=1}^{|\mathbf{d}|} p(d_i\,|\mathbf{r}) \cdot p(\mathbf{r}) \tag{31}$$

Here, two model assumptions have been made: first the conditional probabilities $p(q\,|\mathbf{d}, \mathbf{r})$ are assumed to be independent of $\mathbf{d}$ (cf. Eq.(29)) and secondly, $p(d_i\,|\mathbf{r}, d_1^{i-1})$ shall not depend on the predecessor terms d_1^{i-1} (cf. Eq.(31)).

B. Variants of interpolation

It remains to specify models for the document representations $\mathbf{r} \in \mathbf{R}$ as well as the distributions $p(q\,|\mathbf{r})$, $p(d\,|\mathbf{r})$,

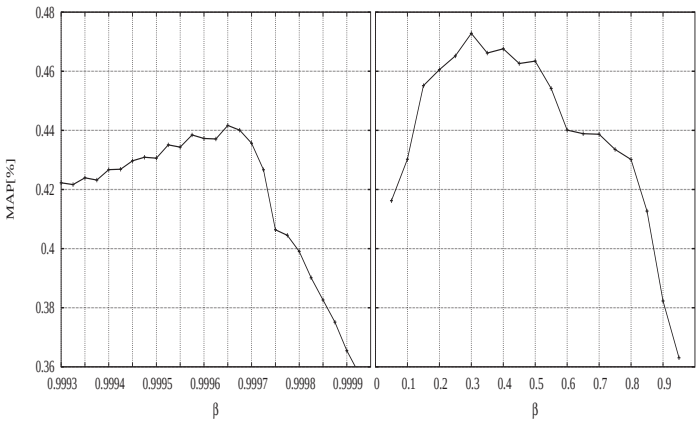

Fig. 3. MAP as a function of the interpolation parameter β according to Eq. (35) (left plot) and Eq. (36) (right plot) with fixed $\alpha = 0.875$ on the reference transcriptions of the TREC-7 SDR task.

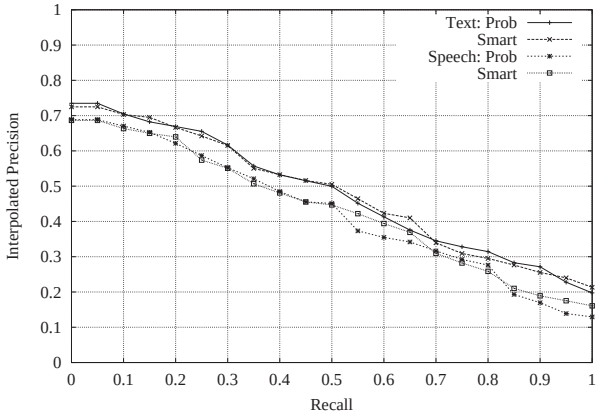

Fig. 4. Interpolated recall-precision graphs for the SMART-2 metric and the new probabilistic approach determined on both the manually transcribed documents (text) and the automatically generated transcriptions (speech) of the TREC-7 spoken document retrieval task.

and $p(\mathbf{r})$. Since we want to distinguish between the event that a query term t is predicted by a representation $\mathbf{r}$ and the event that the term to be predicted is part of a document, $p(q\,|\mathbf{r})$ and $p(d\,|\mathbf{r})$ are modeled differently. In our approach, we identify the set of document representations $\mathbf{R}$ with the histograms over the index terms of the document collection $\mathcal{D}$:

$$n_{\mathbf{r}}(t) \equiv n(t,\mathbf{d}) \qquad n_{\mathbf{r}}(\cdot) \equiv |\mathbf{d}| \tag{32}$$

$$n(t) \equiv \sum_{\mathbf{d}\in\mathcal{D}} n(t,\mathbf{d}) \qquad n(\cdot) \equiv \sum_{\mathbf{d}\in\mathcal{D}} |\mathbf{d}| \tag{33}$$

Thus, we can define the interpolations $p_q(t\,|\mathbf{r})$ and $p_d(t\,|\mathbf{r})$ as models for $p(q\,|\mathbf{r})$ and $p(d\,|\mathbf{r})$:

$$p_q(t\,|\mathbf{r}) := (1-\alpha)\cdot\frac{n_{\mathbf{r}}(t)}{n_{\mathbf{r}}(\cdot)} \;+\; \alpha\cdot\frac{n(t)}{n(\cdot)} \tag{34}$$

$$p_d(t\,|\mathbf{r}) := (1-\beta)\cdot\frac{n_{\mathbf{r}}(t)}{n_{\mathbf{r}}(\cdot)} \;+\; \beta\cdot\frac{n(t)}{n(\cdot)} \tag{35}$$

Since we do not make any assumptions about the a-priori relevance of a document representation, we set up a uniform distribution for $p(\mathbf{r})$. Note that Eq. (35) is an interpolation between the relative counts $n_{\mathbf{r}}(t)/n_{\mathbf{r}}(\cdot)$ and $n(t)/n(\cdot)$. Instead of interpolating between the relative frequencies as in Eq. (35), we can also interpolate between the absolute frequencies:

$$p_d(t\,|\mathbf{r}) := \frac{(1-\beta)\cdot n_{\mathbf{r}}(t) \;+\; \beta\cdot n(t)}{(1-\beta)\cdot n_{\mathbf{r}}(\cdot) \;+\; \beta\cdot n(\cdot)} \tag{36}$$

Both interpolation variants will be discussed in the following section.

C. Experimental results

Experiments were performed on the TREC-7 and the TREC-8 SDR task using both the manually generated transcriptions and the automatically generated transcriptions. As before, all speech recognition outputs were produced using the RWTH LVCSR system for the TREC-7 corpus or taken from the Byblos "Rough 'N Ready" and the Dragon LVCSR system for the TREC-8 corpus.

Due to the small number of test queries for both retrieval tasks, we made use of a leaving-one-out (L-1-O) approach [27, p. 220] in order to estimate the interpolation parameters α and β. Additionally, we added results under unsupervised conditions. i.e. we optimized the smoothing coefficients α and β on TREC-8 queries and corpus and tested on the TREC-7 sets and vice versa. Finally, we carried out a cheating experiment by adjusting the parameters α and β to maximize the MAP on the complete set of test queries. This yields an optimistically upper bound of the possible retrieval effectiveness. All experiments conducted are based on the document representations according to Eq. (32) and Eq. (33), i.e. each document is smoothed with all other documents in the database.

TABLE VI
COMPARISON OF RETRIEVAL EFFECTIVENESS MEASURED IN TERMS OF MAP ON THE TREC-7 SDR TASK FOR THE SMART-2 METRIC AND THE NEW PROBABILISTIC APPROACH PROB. INTERPOLATION WAS PERFORMED ACCORDING TO EQ. (36).

TREC-7	metric		α	β	MAP[%]
text	SMART-2		—	—	46.6
	PROB	"cheating"	0.875	0.300	47.3
		L-1-O	0.742	0.270	**45.8**
		unsupervised	0.950	0.650	42.2
speech (RWTH)	SMART-2		—	—	42.0
	PROB	"cheating"	0.825	0.300	42.0
		L-1-O	0.697	0.257	40.4
		unsupervised	0.875	0.300	**41.6**

TABLE VIII

WERs, TERs, and IERs measured with the Byblos speech recognizer on the Trec-8 corpus for varying preprocessing stages. As before, the substitutions are counted twice for the accumulated error rates of the WER criterion.

		Trec-8			Trec-8 + stop + stem			+ stop + stem, queries only		
	documents	all	relevant	irrelevant	all	relevant	irrelevant	all	relevant	irrelevant
WER[%]	deletions	5.2	6.1	5.1	8.2	7.6	8.2	14.5	11.5	14.7
	insertions	11.3	10.0	11.4	7.6	7.1	7.6	8.6	7.7	8.7
	substitutions	21.9	19.8	22.1	18.2	16.2	18.3	6.2	5.7	6.3
	error rate*	**60.3**	55.6	60.7	52.1	47.1	52.5	35.6	30.7	36.0
TER[%]	deletions	22.3	19.4	22.6	24.2	21.3	24.4	14.2	13.3	14.3
	insertions	29.8	27.2	30.1	24.7	22.5	24.8	20.6	14.5	21.1
	error rate	**52.2**	46.6	52.6	**48.8**	43.8	49.2	**34.8**	27.7	35.4
IER[%]	deletions	16.2	14.9	16.3	17.3	15.4	17.5	10.5	8.8	10.6
	insertions	18.9	17.0	19.1	17.4	15.9	17.5	11.8	9.2	12.0
	error rate	35.1	31.9	35.4	34.7	31.4	35.0	**22.3**	**18.0**	22.7

TABLE IX

WERs, TERs, and IERs measured with the Dragon speech recognizer on the Trec-8 corpus for varying preprocessing stages. As before, the substitutions are counted twice for the accumulated error rates of the WER criterion.

		Trec-8			Trec-8 + stop + stem			+ stop + stem, queries only		
	documents	all	relevant	irrelevant	all	relevant	irrelevant	all	relevant	irrelevant
WER[%]	deletions	6.5	6.9	6.5	8.9	7.4	9.1	15.6	11.5	15.9
	insertions	12.7	11.2	12.9	8.0	7.5	8.0	9.4	8.3	9.5
	substitutions	21.0	18.5	21.2	17.7	15.6	17.9	6.2	5.3	6.2
	error rate*	**61.3**	55.0	61.8	52.3	46.2	52.8	37.3	30.3	37.9
TER[%]	deletions	22.8	19.2	23.1	24.5	20.7	24.8	14.6	13.2	14.8
	insertions	24.7	22.4	24.9	22.0	14.6	22.7	29.8	27.2	30.1
	error rate	**53.2**	46.6	53.8	**49.2**	43.0	49.7	**36.7**	27.8	37.4
IER[%]	deletions	17.0	15.0	17.1	17.9	15.2	18.1	11.0	9.3	11.2
	insertions	19.7	17.8	19.9	17.6	16.3	17.7	12.4	9.4	12.6
	error rate	36.7	32.7	37.0	35.5	31.5	35.8	**23.4**	**18.7**	23.8

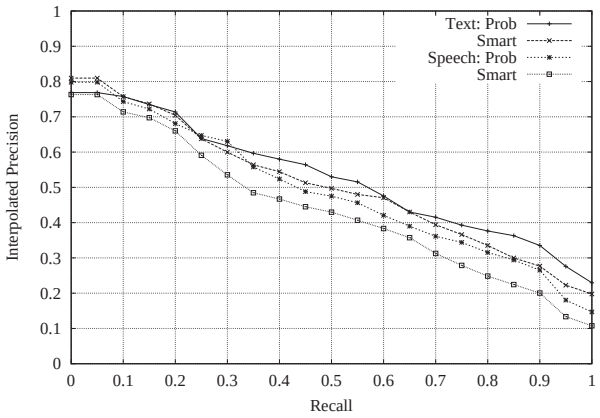

Fig. 5. Interpolated recall-precision graphs for the Smart-2 metric and the new probabilistic approach determined on both the manually transcribed documents (text) and the automatically generated transcriptions (speech) of the Trec-8 spoken document retrieval task.

In a first experiment, the interpolation parameter α was estimated. Fig. 2 shows the MAP as a function of the interpolation parameter α with fixed β on the reference transcriptions of the Trec-7 corpus. Using the L-1-0 estimation scheme, the best value for α was found to be 0.742 which has to be compared with a globally optimal value of 0.875, i.e. the cheating experiment without L-1-O. The interpolation parameter β was adjusted in a similar way. Using the interpolation scheme according to Eq. (35), the retrieval effectiveness on both tasks is maximum for

TABLE VII

Comparison of retrieval effectiveness measured in terms of mean average precision (MAP) on the Trec-8 spoken document retrieval task for the Smart-2 metric and the new probabilistic approach Prob. Interpolation was performed according to Eq. (36).

Trec-8	metric		α	β	MAP[%]
text	Smart-2		—	—	49.6
	Prob	"cheating"	0.950	0.650	52.7
		L-1-O	0.947	0.646	**51.3**
		unsupervised	0.875	0.300	49.9
speech (Byblos)	Smart-2		—	—	43.1
	Prob	"cheating"	0.875	0.300	47.3
		L-1-O	0.801	0.287	44.4
		unsupervised	0.825	0.300	**47.2**
speech (Dragon)	Smart-2		—	—	42.1
	Prob	"cheating"	0.875	0.300	45.6
		L-1-O	0.875	0.307	44.1
		unsupervised	0.825	0.300	**45.2**

values of β that are very close to 1. This effect is caused by singletons, i.e. index terms that occur once only in the whole document collection. Since the magnitude of the ratio of both denominators in Eq. (35) is approximately

$$\frac{n_{\mathbf{r}}(\cdot)}{n(\cdot)} \approx \frac{1}{D}$$

the optimal value for β should be found in the range of $1 - 1/D$, assuming that singletons are the most important features in order to filter out a relevant document. In fact, using $\beta = 1 - 1/D$ exactly meets the optimal value of 0.99965 on the TREC-7 corpus and 0.99995 on the TREC-8 retrieval task.

However, since the interpolation according to Eq. (35) runs the risk of becoming numerically unstable (especially for very large document collections), we investigated an alternative smoothing scheme that interpolates between absolute counts instead of relative counts (cf. Eq. (36)). Fig. 3 depicts the MAP as a function of the interpolation parameter β for both interpolation methods on the reference transcriptions of the TREC-7 SDR task. Since the interpolation scheme according to Eq. (36) proved to be numerically stable and achieved slightly better results, it was used for all further experiments. Table VI shows the obtained retrieval effectiveness for the new probabilistic approach on the TREC-7 SDR task. Using L-1-O, the retrieval performance of the new proposed method lies within the magnitude of the SMART-2 metric, i.e. we obtained a MAP of 45.8% on manually transcribed data, which must be compared with 46.6% using the SMART-2 retrieval metric. Using automatically generated transcriptions we achieved a MAP of 40.4% which is close to the performance of the SMART-2 metric. A further improvement gain could be obtained under unsupervised conditions. Using the optimal parameter setting of the TREC-8 corpus for the TREC-7 task yielded a MAP of 41.6%. Fig. 4 shows the recall-precision graphs for both SMART-2 and the new probabilistic approach.

The same applies to the results obtained on the TREC-8 SDR task (cf. Table VII). Here, the new probabilistic approach even outperformed the SMART-2 retrieval metric. Thus, we obtained a MAP of 51.3% on the manually transcribed data in comparison with 49.6% for the SMART-2 metric. This improvement over SMART-2 is also obtained on recognized transcriptions even though the improvement is smaller. Thus, we achieved a MAP of 44.4% on the automatically generated transcriptions produced with the Byblos speech recognizer, which is an improvement of 3% relative compared to the SMART-2 metric, and 44.1% MAP using the Dragon speech recognition outputs, which is an improvement of 5% relative. Similar to the results obtained on the TREC-7 corpus the unsupervised experiments conducted on the automatically generated transcriptions of the TREC-8 task showed a further performance gain between 1% and 2% absolute. Fig. 5 shows the recall-precision graphs for SMART-2 and the probabilistic approach.

V. CONCLUSION

In this paper, we presented a detailed analysis on the effect of recognition errors on retrieval performance. Since retrieval performance is only little affected by recognition errors, we investigated two alternative error measures, namely the *term error rate* and the *indicator error rate* that turned out to be more suitable in order to describe the quality of automatically generated transcriptions for retrieval applications. Experiments carried out on the TREC-7 and TREC-8 spoken document retrieval task revealed a better correlation between the obtained retrieval effectiveness and the proposed error measures. Baseline results were produced using a new query expansion method.

In the second part of this article, we presented a new probabilistic approach to spoken document retrieval based on interpolations between document specific term histograms and a global term histogram that is pooled over all documents. To this purpose, the set of documents was mapped onto a set of document representations. These document representations were identified with document specific histograms and can be interpreted as a kind of nearest neighbor concept. Two smoothing schemes were discussed and investigated. Experiments performed on the TREC-7 and the TREC-8 spoken document retrieval task showed comparable or even better results for the new probabilistic approach than an enhanced version of the SMART-2 retrieval metric. In addition, the new probabilistic approach turned out to be robust towards recognition errors.

REFERENCES

[1] W. Liggett and W. Fisher, "Insights from the broadcast news benchmark tests," in *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Va, USA, February 1998, pp. 16–22.

[2] J. S. Garofolo, E. M. Voorhees, C. G. P. Auzanne, V. M. Stanford, and B. A. Lund, "1998 TREC-7 spoken document retrieval track: overview and results," in *Proc. 7th Text REtrieval Conference (TREC 7)*, Gaithersburg, Md, USA, November 1998, vol. 500-242 of NIST Special Publication, pp. 79–89.

[3] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proc. 8th Text REtrieval Conference (TREC 8)*, Gaithersburg, Md, USA, 1999, vol. 500-246 of NIST Special Publication, pp. 107–130.

[4] M. F. Porter, "An algorithm for suffix stripping," July 1980, Programm.

[5] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," in *Proc. of 4th Text REtrieval Conference (TREC-4)*, D. K. Harman, Ed., National Institute of Standards and Technology, Gaithersburg, Md, USA, October 1996, pp. 73–96.

[6] A. Singhal, J. Choi, D. Hindle, D. D. Lewis, and F. C. N. Pereira, "AT&T at TREC-7," in *Proc. 7th Text REtrieval Conference (TREC-7)*, Gaithersburg, Md, USA, November 1998, vol. 500-242 of NIST Special Publication, pp. 239–252.

[7] J. Choi, D. Hindle, J. Hirschberg, I. Magrin-Changnolleau, C. Nakatani, F. Pereira, A. Singhal, and S. Whittaker, "An overview of the AT&T spoken document retrieval," in *Proc.*

1998 DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Va, USA, February 1998, pp. 182–188.

[8] J. J. Rocchio, "Relevance feedback in information retrieval," in *The SMART Retrieval System – Experiments in Automatic Document Processing*, Englewood Cliffs, NJ, USA, 1971, pp. 313–323, Prentice Hall.

[9] W. Cohen and Y. Singer, "Context-sensitive learning methods for text categorization," in *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, August 1996, pp. 307–315.

[10] R. Schapire, Y. Singer, and A. Singhal, "Boosting and rocchio apllied to text filtering," in *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998, pp. 215–223.

[11] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," *ACM Transactions on Information Systems*, vol. 18, no. 1, pp. 79–112, January 2000.

[12] S. Kanthak, A. Sixtus, S. Molau, R. Schlüter, and H. Ney, "Fast search for large vocabulary speech recognition," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed., pp. 63–78. Springer-Verlag, Berlin, Germany, 2000.

[13] S. Ortmanns, A. Eiden, and H. Ney, "Improved lexical tree search for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, Wash, USA, May 1998, vol. 2, pp. 817–820.

[14] Peter Schäuble, *Multimedia Information Retrieval*, Kluwer Academic, Boston, Mass, USA, 1997.

[15] F. Kubala, S. Colbath, D. Liu, A. Srivastava, and J. Makhoul, "Integrated technologies for indexing spoken language," *Communications of the ACM*, vol. 43, no. 2, pp. 48, February 2000.

[16] S. Wegmann, P. Zhan, I. Carp, M. Newman, J. P. Yameon, and L. Gillick, "Dragon systems' 1998 broadcast news transcription system," in *Proc. 1999 DARPA Broadcast News Workshop*, Herndon, Va, USA, February–March 1999, pp. 277–280.

[17] S.E. Johnson, P. Jourlin, G.L. Moore, K. Spärck Jones, and P.C. Woodland, "Spoken document retrieval for TREC-7 at Cambridge University," in *Proc 7th Text REtrieval Conference (TREC-7)*, Gaithersburg, Md, USA, November 1999, vol. 500-242 of NIST Special Publication, pp. 191–200.

[18] E. Mittendorf and P. Schäuble, "Measuring the effects of data corruption on information retrieval," in *Proc. Symposium on Document Analysis and Information Retrieval (SDAIR 96)*, Las Vegas, Nev, USA, April 1996, pp. 179–189.

[19] N. Fuhr and C. Buckley, "A probabilistic learning approach for document indexing," *ACM Transcations on Information Systems*, vol. 9, no. 3, pp. 223–248, 1991.

[20] J. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998, pp. 275–281.

[21] A. Berger and J. D. Lafferty, "Information retrieval as statistical translation," in *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, Calif, USA, August 1999, pp. 222–229.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.

[23] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[24] D. R. H. Miller, T. Leek, and R. M. Schwartz, "BBN at TREC7: Using hidden Markov models for information retrieval," in *Proc. 7th Text REtrieval Conference (TREC-7)*, Gaithersburg, MD, USA, November 1999, vol. 500-242 of NIST Special Publication, pp. 133–142.

[25] J. L. Gauvain, Y. de Kercadio, L. F. Lamel, and G. Adda, "The LIMSI SDR system for TREC-8," in *Proc. 8th Text REtrieval Conference TREC-8*, Gaithersburg, Md, USA, November 1999, pp. 405–412.

[26] H. J. Viechtbauer, "Vergleich heuristischer und statistischer Verfahren im Information Retrieval," Diploma thesis, Lehrstuhl für Informatik VI, Computer Science Department, RWTH Aachen, University of Technology, Aachen, Germany, September 2001.

[27] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, San Diego, Calif, USA, 2nd edition, 1990.

Wolfgang Macherey received the Diploma degree with honor in computer science in 1999 from Aachen University of Technology, Germany.
Since 1999, he has been a Research Assistant with the Department of Computer Science of Aachen University of Technology. From July to September 2002, he was a summer student at IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests are in large-vocabulary speech recognition, acoustic modeling with the focus on discriminative training and affine feature space transformations as well as in information retrieval.

Hans Jörg Viechtbauer received the Diploma degree in computer science in 2002 from Aachen University of Technology, Germany.
From July 2000 to February 2002 he was a research supplemental at the Department of Computer Science of Aachen University of Technology. Since July 2002, he has been with RecomMind GmbH, Rheinbach, Germany. His research interests are in information retrieval, speech recognition, language modeling, and pattern recognition.

Hermann Ney received the Diploma degree in physics in 1977 from Göttingen University, Germany, and the Dr.-Ing. degree in electrical engineering in 1982 from Braunschweig University of Technology, Germany.
He has been working in the field of speech recognition, natural language processing, and stochastic modeling for more than 20 years and has authored and co-authored more than 200 papers. In 1977, he joined Philips Research in Germany. In 1985, he was appointed Department Head. All of his career at Philips was in research and advanced development of basic technology for pattern recognition, speech recognition and spoken language systems. From October 1988 to October 1989, he was a Visiting Scientist at Bell Laboratories, Murray Hill, NJ. In July 1993, he joined the Computer Science Department of Aachen University of Technology as a Full Professor. His responsibilities include planning, directing and carrying out research for national, European and industrial sponsors and supervising PhD students. He has been and is a peer reviewer for a number of major scientific journals. He is on the editorial board of several major scientific journals.
From 1992 to 1998, he was on the Executive Board of the German section of the IEEE. For the term 1997–2000, he was a member of the Speech Technical Committee of the IEEE.