Using String Comparison in Context for Improved Relevance Feedback in Different Text Media

Adenike M. Lam-Adesina and Gareth J. F. Jones

Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland
email: {adenike,gjones}@computing.dcu.ie

Abstract. Query expansion is a long standing relevance feedback technique for improving the effectiveness of information retrieval systems. Previous investigations have shown it to be generally effective for electronic text, to give proportionally better improvement for automatic transcriptions of spoken documents, and to be at best of questionable utility for optical character recognized scanned text documents. We introduce two corpus-based methods based on using a string-edit distance measure in context to automatically detect and correct transcription errors. One method operates at query-time and requires no modification of the document index file, and the other at index-time and operates using the standard query-time expansion process. Experimental investigations show these methods to produce improvements in relevance feedback for all three media types, but most significantly mean that relevance feedback can now successfully be applied to scanned text documents.

1 Introduction

Query expansion within relevance feedback (RF) has been shown to improve effectiveness for many information retrieval (IR) tasks, However, its performance varies for different text media for the same retrieval task. Performance differences arise from the indexing errors associated with the individual media. In this study we are concerned with improving the effectiveness of relevance feedback for a common retrieval task for the following text media: standard typed electronic text, transcriptions of spoken data created using automatic speech recognition (ASR), and transcriptions of scanned paper text documents generated using optical character recognition (OCR). While the accuracy of automatically generated digital document transcriptions continues to increase with advances in recognition technologies, the error levels are likely to remain sufficient to adversely affect relevance feedback effectiveness for the foreseeable future. Current transcription technologies achieve good performance on tasks such as read speech and recently printed texts, but still have very significant error levels for more challenging tasks such as conversational speech in noisy environments and nth generation photocopies or hand written texts. It can be observed that even the level of typographical errors found in published electronic texts can be sufficient to be detrimental to relevance feedback [1].

Relevance feedback using query expansion has previously been explored for all three media for both true relevance feedback using user-entered relevance judgements

and pseudo relevance feedback (PRF) where the top ranked documents in the initial retrieval run are assumed to be relevant. The general conclusions for these studies are as follows: on average relevance feedback improves retrieval precision for typed text retrieval (TR), gives a proportionally greater improvement for spoken document retrieval (SDR) than text retrieval, but is not generally effective for document image retrieval (DIR). In fact relevance feedback can often reduce average performance for DIR.

In this paper we describe two string-based methods to improve relevance feedback performance for these text media. One operates at query-time and can be applied to existing document search collections without re-indexing. The other operates at indexing time and requires no modification of the standard query expansion process for relevance feedback at query-time. Both techniques apply a string-edit distance measure in context to identify likely misspellings or incorrectly transcribed valid words, and then seek to correct them from within the document collection. Both methods produce effective relevance feedback for DIR, and small improvements in relevance feedback for text retrieval, and the index-time method an improvement in SDR.

This paper focuses on the retrieval effectiveness of PRF in terms of standard precision and recall metrics. This is an experimental investigation and, as such, issues of computational efficiency of the implementation of the relevance feedback process are beyond the scope of this study.

The remainder of this paper is organised as follows: Section 2 gives a short review of relevant existing research, Section 3 outlines details of the Okapi BM25 information retrieval model used in this investigation, Section 4 summarizes the details of the test collection, Section 5 describes our extended relevance feedback methods and results for experiments using these techniques, and finally Section 6 concludes the paper.

2 Relevant Existing Research in Relevance Feedback

This section gives a brief summary of existing work in relevance feedback relevant to this paper. While relevance feedback has been studied for many years for different tasks, many recent investigations have taken place within tasks at the TREC workshops [2], including the main ad hoc search tasks and tasks focusing on other media. Another notable activity exploring relevance feedback was the Reliable Information Access (RIA) workshop in 2003 [3]. Within TREC, relevance feedback studies have mainly explored PRF, with results generally indicating that across a topic set PRF on average produces improvements in the standard TREC evaluation metrics.

Existing SDR studies of relevance feedback have again focused primarily on PRF within TREC SDR tasks [4]. Results here in general indicate that PRF is very effective for SDR with collections of automatically transcribed broadcast news [5]. These results are confirmed for a very different retrieval task of unstructured oral testimonies in the speech retrieval task introduced at CLEF 2005 [6].

The main results for DIR are again from TREC, this time in the confusion track [7]. Although participants explored a range of relevance feedback methods, the results were inconclusive since this was a single known-item search task. The absence of exhaustive document relevance information meant that it was not possible to study the effects of relevance feedback techniques thoroughly. Much more extensive evaluation of DIR has

been carried out at the University of Nevada at Las Vegas [8]. Results from these studies were again inconclusive, but suggested that relevance feedback is much less reliable for DIR than text retrieval and SDR.

In an earlier study of PRF for a parallel document collection for text retrieval, SDR and DIR we showed good performance for text retrieval, better relative performance for SDR, but a significant reduction in average precision and recall for DIR [9]. This result motivated us to investigate both the reasons for the ineffectiveness of PRF for DIR, and more generally to seek to understand the reasons for the variations in PRF effectiveness for different text media. In a previous study [1], we showed that the principal problem for PRF in DIR is the presence in the collection of high numbers of very rare index terms which are in fact character corrupted versions of standard words. While correctly spelled versions of these words have standard expected word frequencies within the collection as a whole.

In this paper we describe two techniques which address PRF problems for DIR and illustrate how these can also be effective for text retrieval and SDR.

3 Information Retrieval and Relevance Feedback Methods

The basis of our experimental setup is the City University research distribution version of the Okapi system [10]. The Okapi retrieval model has been shown to be very effective in many comparative evaluation exercises in recent years at TREC and elsewhere. The retrieval strategy adopted in this investigation follows standard practice for best-match ranked retrieval. The documents and search topics are first processed to remove common stop words from a list of around 260 words, suffix stripped using the Okapi implementation of Porter stemming to encourage matching of different word forms, and terms are further indexed using a small set of synonyms.

3.1 Term Weighting

Following preprocessing document terms are weighted using the Okapi BM25 weight [10]. The BM25 weight for a term is calculated as follows,

$$cw(i,j) = cfw(i) \times \frac{tf(i,j) \times (k_1 + 1)}{k_1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where $cw(i,j)$ represents the weight of term i in document j, $cfw(i) = log((N - n(i) + 0.5)/(n(i) + 0.5))$, $n(i)$ is the total number of documents containing term i, and N is the total number of documents in the collection, $tf(i,j)$ is the within document term frequency, and $ndl(j) = dl(j)/\text{Av.}dl$ is the normalized document length where $dl(j)$ is the length of j. k_1 and b are empirically selected tuning constants for a particular collection. The matching score for each document is computed by summing the weights of terms appearing in the query and the document.

3.2 Relevance Feedback

In the standard Okapi approach potential expansion terms are ranked using the Robertson's offer weight ($ow(i)$) [10], defined as,

$$ow(i) = r(i) \times rw(i) \tag{1}$$

where $r(i)$ is the number of relevant documents containing term i, and $rw(i)$ is the standard Robertson/Sparck Jones relevance weight [10] defined as,

$$rw(i) = \log \frac{(r(i) + 0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)}$$

where $n(i)$ and N have the same definitions as before and R is the total number of relevant documents for this query. The top ranking terms are then added to the original query. Term reweighting for relevance feedback is carried out by replacing $cfw(i)$ with $rw(i)$ in the BM25 weight. In this study we explore only query expansion since we generally observe this to be the dominant factor in relevance feedback.

Selection of expansion terms from whole documents can result in query drift if terms associated with non-relevant material are selected. In this study we adopt our sentence-based query-biased summary technique described in [11]. In this procedure potential expansion terms are selected from the query-biased summary of each potentially relevant document. This method has been shown to reduce the possibility of query drift in previous studies. Potential expansion terms are selected from the top R_1 documents assumed relevant, but the $ow(i)$ is calculated using a separate larger R value, since we find this to give more effective $ow(i)$'s.

4 Test Collections

The experimental investigation was carried out using a parallel research collection of text, spoken and image documents adapted from the TREC-8 SDR task [4]. The original SDR test collection consisted of the documents, search requests and relevant documents for each request. For our investigation we used a parallel document image collection consisting of scanned images generated from manual transcriptions of the audio data. The TREC-8 SDR collection is based on the English broadcast news portion of the TDT-2 News Corpus. The standard SDR collection of text and spoken document sets is augmented by forming a corresponding scanned document collection. The scanned document collection is based on the 21,759 "NEWS" stories in TDT-2 Version 3 (December 1999).

4.1 TDT-2 Document Set

The TREC-8 SDR portion of the TDT-2 News Corpus covers a period of 5 months from February to June 1998. The collection consists of 30 minute news broadcasts from CNN, ABC, PRI and VOA. Each broadcast is manually segmented into a number of news stories with unique identifers which form the basic document unit of the corpus.

An individual news story was defined as containing two or more declarative statements about a single event. Other miscellaneous data items, e.g. commercials, were excluded from the data set. The collection contains a total of 21,759 stories with an average length of 180 words totalling about 385 hours of audio data.

Text Collection There is no high-quality human reference transcription available for TDT-2 - only "closed-caption" quality transcriptions for the television sources and rough manual transcriptions for the radio sources made by commercial transcription services. A detailed manual transcription of a randomly selected 10 hour subset was carried out by the corpus developers to enable speech recognition accuracy to be evaluated. The television closed-caption sources (CNN, ABC) were found to have a Word Error Rate of approximately 14.5% and radio sources (PRI, VOA) to have a Word Error Rate of around 7.5%. The manual transcriptions are used as the document source for the scanned document collection used in this study.

Spoken Document Collection The Spoken Document transcriptions used in our experiments are taken from the TDT-2 version 3 CD-ROMs. The transcription set used is designated `as1` on this release and was generated by NIST using the BBN BYBLOS Rough'N'Ready transcription system using a dynamically updated rolling language model. Full details of this recognition system are contained in [12]. This transcription was designated "B2" in the official NIST TREC-8 SDR documentation. The recognition Word Error Rate on a 10 hour subset of the data was reported by the developers to be 26.7%.

Scanned Document Collection The printed version of the collection is formatted as hardcopy similar in style to newspaper clippings. To simulate the differences in formatting of stories from different newspaper sources, each story was printed in one of four fonts: *Times*, *Pandora*, *Computer Modern* and *San serif*. The stories were divided roughly equally between these font types with material from each source assigned to each one on a sequential basis. The stories were printed in one of three font sizes in single columns in one of six widths. Column width and font size were assigned sequentially from the beginning of each broadcast. The stories were printed using an Epson EPL-N4000 laser printer. In order to explore retrieval behaviour with a more errorful transcription than would naturally result from a printing of this quality, OCR transcription was performed with suboptimal system settings. All documents were scanned using an HPScanJet ADF at 200 dpi in Black & White at a threshold of 100. OCR was carried out using Page Keeper Standard Version 3.0 (OCR Engine Version 271) (SR3). Full details of the collection design are contained in [13].

4.2 TREC-8 SDR Test Collection

The TREC-8 SDR retrieval test collection contains a set of 50 search topics and corresponding relevance assessments. The goal in creating the topics was to devise topics with a few (but not too many) relevant documents in the collection to appropriately challenge test retrieval systems. Retrieval runs submitted by the TREC-8 SDR participants were used to form document pools for manual relevance assessment. The average topic length was 13.7 words and the mean number of relevant documents for each topic was 36.4 [4].

Table 1. Baseline and standard summary-based feedback results for TR, SDR and DIR.

Media	TR				SDR				DIR			
	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
Baseline	0.551	0.354	0.468	1608	0.496	0.321	0.406	1502	0.557	0.352	0.454	1581
Fbk 5	0.580	0.392	0.506	1639	0.500	0.346	0.423	1514	0.574	0.380	0.498	1578
chg bl. (%)	+5.3	+10.7	+8.1	+31	+0.8	+7.8	+4.2	+12	+3.1	+7.9	+9.7	-3
Fbk 20	0.598	0.396	0.514	1631	0.553	0.361	0.459	1532	0.539	0.352	0.440	1385
chg bl. (%)	+8.5	+11.9	+9.8	+23	+11.5	+12.5	+13.1	+30	-4.1	-0	-3.1	-196

5 Investigation of Relevance Feedback for Different Text Media

This section reports our investigation of query expansion for different text media. Results are shown for standard PRF methods, and our new techniques for enhancing the effectiveness of PRF. Retrieval metrics reported are precision at 10 and 30 document cutoff, standard TREC average precision (AvP) and the total number of relevant documents retrieved (RelRet). The total number of relevant documents retrieved for each run can be compared for Recall to the total number of relevant documents available across all topic statements in the TREC-8 SDR test set of 1818. Percentage change relative to a no PRF baseline is shown for precision measures and absolute change for the number of relevant documents retrieved.

The BM25 values were set empirically as $k_1 = 1.4$ and $b = 0.6$ using the baseline retrieval system with the text document collection without PRF to optimise AvP. The parameters for the summary-based PRF were set as follows. Summaries were based on the most significant 6 sentences, the top 5 ranked documents are the source of potential feedback terms (R_1), and the top 20 documents assumed relevant for computation of $ow(i)$ for term selection (R). The weight of the original query terms in each case was multiplied by 1.5 relative to the expansion terms, since the original terms have been chosen by the searcher themself. These values were again selected to optimise AvP on the text collection

5.1 Baseline and Standard PRF Results

Table 1 shows baseline retrieval results without feedback (Baseline) and PRF results adding 5 (Fbk 5) and 20 (Fbk 20) expansion terms for text retrieval, SDR and DIR, with their changes from the baseline. From Table 1 it can be seen that there is a reduction in both baseline AvP and RelRet for SDR and a smaller one for DIR compared to TR. For PRF results, performance in terms of both precision and RelRet improves in all cases for TR and SDR. For DIR, PRF improves for 5 expansion terms, but AvP decreases by -3.1% and RelRet by -196 for 20 expansion terms.

In previous work [1] we demonstrated that the reduction in PRF performance for DIR is due to selection of some expansion terms with very low $n(i)$ values which are misrecognized versions of more common terms corrupted at the character level. We could of course try to correct these errors in post-processimg with a dictionary, however existing work [14] indicates that this can introduce more problems for information retrieval due to false substitutions than it solves. We thus do not explore dictionary-based substitution methods.

5.2 Improving PRF by String-Based Compensation for Transcription Errors

In previous work [1], we demonstrated that a simple filtering of terms with low $n(i)$ values partially addresses the problems with PRF for DIR associated with spelling mistakes illustrated in Table 1. However, the optimal value of $n(i)$ for filtering may be sensitive to the statistics of individual collections. Additionally, there are two notable problems with this very basic approach. First, correctly transcribed rare words that would actually be good expansion terms will be deleted along with the incorrectly transcribed ones, and thus not be available as potential expansion terms. Second, many incorrectly transcribed, apparently rare, words can be recognized manually as corrupted versions of correct terms appearing in assumed relevant documents. These variant forms are obvious to a human reader based on string similarity and the linguistic context in which they are found. Further, in such cases the $ow(i)$ values of the correctly transcribed terms will often be wrong since $r(i)$ will be underestimated when there is no other occurrence of i in a document within which it is incorrectly transcribed. Terms of this type will often be in high ranked documents for a query for which the terms are important. This is potentially a significant problem leading to distortion in the ranking of the $ow(i)$ ordered list compared to the one that would be formed without spelling errors in the documents, consequential reduction in the likelihood of choosing the best expansion terms, and thus potentially reduction in the possible effectiveness of relevance feedback. Spelling mistakes in text documents can also on occasion lead to similar problems for PRF in text retrieval which are not visible when looking across averaged results such as those shown in Table 1. While PRF works effectively for SDR, and the fixed vocabulary of automatic speech recognition systems used to generate automatic transcriptions means that misspellings of this type are not possible, the high overall Word Error Rate does affect retrieval effectiveness and means that there is scope to improve performance beyond that seen in Table 1.

Problems of eliminating good potential expansion terms and inaccurate estimates of $ow(i)$ can be overcome by identifying mistranscribed words within (assumed) relevant documents and combining them with correct words. In this section we introduce two methods for doing this. The first is a query-time technique that can be applied to existing indexed collections. While this method is found to be effective, it imposes an additional search time computational load. The second technique operates at index-time and imposes no additional search time cost.

Both procedures are based on a string comparison algorithm which computes an "edit distance" between two strings giving the minimum number of changes required to convert one string to the other [15]. These algorithms can make mistakes, sometimes merging words that are not related. However, within the constrained context of a small number of documents assumed to be relevant to a search query, often similar character strings really are the same word, leading to only a small number of false merges. This hypothesis is used as the basis of our correction techniques.

Query-Time Expansion Term Combination In the query-time procedure the edit distance is computed between all terms within the top 5 ranked summaries used for PRF. Words within a preset edit distance are merged with the one with the larger $n(i)$ value assumed to be the correct. The $r(i)$ values of merged words are added, and the

Table 2. Results using string-comparison term merging at query-time.

Media	TR				SDR				DIR			
MaxEd	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
1	0.598	0.384	0.519	1614	0.551	0.362	0.459	1531	0.576	0.380	0.489	1593
chg bl. (%)	+8.5	+8.5	+10.9	+6	+11.1	+12.8	+13.1	+29	+3.4	+8.0	+7.7	+12
2	0.608	0.387	0.524	1614	0.553	0.362	0.456	1530	0.582	0.380	0.492	1581
chg.bl. (%)	+10.3	+9.3	+12.0	+6	+11.5	+12.8	+12.3	+28	+4.5	+8.0	+8.4	+0
3	0.610	0.399	0.528	1624	0.553	0.374	0.465	1541	0.596	0.385	0.505	1610
chg. bl. (%)	+10.7	+12.7	+12.8	+16	+11.5	+16.5	+14.5	+39	+7.0	+9.4	+11.2	+29
4	0.604	0.399	0.521	1639	0.549	0.363	0.454	1533	0.588	0.388	0.508	1616
chg. bl. (%)	+9.6	+12.7	+11.3	+31	+10.1	+13.1	+11.8	+31	+5.6	+10.2	+11.9	+35
5	0.598	0.393	0.523	1616	0.537	0.369	0.450	1552	0.592	0.386	0.507	1607
chg. bl. (%)	+8.5	+11.6	+11.8	+8	+8.3	+15.5	+10.8	+50	+6.3	+9.7	+11.7	+26

combined $n(i)$ value is taken as that of the larger value. The reduced set of potential expansion terms is then ranked by the $ow(i)$ computed using the merged $r(i)$ values.

Table 2 shows the result of using this merging approach with 20 expansion terms for maximum edit distance values of 1, 2, 3, 4 and 5, for TR, SDR and DIR. From Table 2 it can be seen that AvP is improved for both TR and DIR compared to the results shown in Table 1. For DIR performance clearly improves as the maximum distance is increased to 4 characters. There is little variation between results for maximum allowed edit distance values of 3, 4 and 5, suggesting that using a value of 4 will give good average stability across different queries for this collection. For TR the best maximum edit distance is 2 or 3, although any value above 1 gives very similar results. The technique does not appear to be effective for SDR; there is one AvP result above those in Table 1, but overall there is no trend indicating improvement. Analysis of $n(i)$ values in the speech documents shows that very few terms in the automatic transcription have low $n(i)$ values and occasional misspellings are not possible, and thus as observed the method has little scope for impact on PRF for SDR.

The success of this technique for TR and DIR can be attributed to the elimination of highly weighted rare misrecognized terms from the feedback terms for two reasons. First, the rank of non-relevant documents in the assumed relevant set which contain these terms is not now promoted by addition of these highly weighted terms to the search query. The rank of non-relevant documents may still be promoted due to the presence of other expansion terms, but this is a general drawback of query expansion in PRF for all media types. Second, in addition to their presence in the assumed relevant document set, although rare, if their $n(i)$ value > 1 these individual misrecognized terms can also occur in other documents, effectively with a random distribution. These other documents containing incorrect terms may include some or none of the original query terms, but when the query is expanded to include the highly weighted errorful terms, the matching score of the documents containing them can increase dramatically relative to other documents. While these documents may be relevant to the search request, it is most likely that they will often not be relevant.

Overall then for TR and DIR the merging technique gives better estimation of $ow(i)$ due to more accurately calculating $r(i)$, and prevents problems of over promotion of documents containing errorful terms. While effective and not requiring re-indexing of

Table 3. Baseline and PRF results for indexing-time term combination ($e < 4$, $m > 4$, $R_1 = 5$).

Media	TR				SDR				DIR			
	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
Baseline	0.451	0.365	0.476	1601	0.492	0.328	0.406	1487	0.559	0.359	0.464	1543
Fbk 5	0.596	0.386	0.508	1598	0.533	0.359	0.434	1520	0.598	0.378	0.499	1540
chg bl. (%)	+10.2	+5.8	+6.7	-3	+8.3	+9.5	+6.9	+33	+6.9	+5.3	+7.5	-3
Fbk 20	0.602	0.406	0.519	1617	0.533	0.365	0.458	1517	0.590	0.394	0.499	1600
chg. bl. (%)	+11.3	+11.2	+9.0	+16	+8.3	+11.3	+12.8	+30	+5.5	+9.7	+7.5	+57

Table 4. Baseline and PRF results for indexing-time term combination ($e < 4$, $m > 1$, $R_1 = 10$).

Media	TR				SDR				DIR			
	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
Baseline	0.543	0.371	0.467	1568	0.498	0.335	0.414	1492	0.549	0.362	0.457	1539
Fbk 5	0.565	0.382	0.489	1573	0.535	0.366	0.451	1523	0.584	0.383	0.484	1558
chg bl. (%)	+4.1	+3.0	+4.7	1614	+7.4	+9.3	+8.9	+31	+6.4	+5.8	+5.9	+19
Fbk 20	0.588	0.412	0.525	1610	0.543	0.383	0.468	1549	0.590	0.399	0.503	1580
chg. bl. (%)	+8.3	+11.1	+12.4	+42	+9.0	+14.3	+13.0	+57	+7.5	+10.2	+10.1	+41

Table 5. Baseline and PRF results for indexing-time term combination ($e < 4$, $m > 4$, $R_1 = 10$).

Media	TR				SDR				DIR			
	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
Baseline	0.539	0.369	0.472	1555	0.494	0.335	0.412	1487	0.539	0.378	0.464	1552
Fbk 5	0.600	0.391	0.510	1555	0.553	0.366	0.467	1493	0.569	0.378	0.495	1516
chg bl. (%)	+11.3	+6.0	+8.1	+0	+11.9	+8.5	+13.3	+6	+5.6	+0.0	+6.7	-36
Fbk 20	0.592	0.399	0.519	1539	0.561	0.389	0.483	1538	0.563	0.391	0.500	1591
chg. bl. (%)	+9.8	+8.1	+10.0	-16	+13.5	+16.1	+17.2	+51	+4.5	+3.4	+4.5	+39

the document collection, this method imposes a potentially significant computational load at query-time. In the next section we describe an index-time correction method using string-comparison in context which enables standard PRF methods to be used without modification.

Index-Time Combination for Term Correction In addition to imposing a query-time computational load, having identified mistakes the search time technique cannot actually correct the mistakes in the documents. Thus incorrectly transcribed words in documents will still not match with the expanded query in the feedback retrieval run, there is no modification to term weights in the feedback retrieval run (for example based on correction of $tf(i, j)$ values), and the merging must be carried out each time a word appears in a new query.

In most cases when an incorrect word occurs in a document, we observe that it is often the case that the word appears correctly in other documents covering similar topics. We exploit this observation to correct mistranscribed words by using correctly transcribed ones in similar contexts. This procedure operates as follows.

All individual documents are converted into queries. Each document query is then used to query the complete original document collection. It is expected that the query

(document) will retrieve itself in rank position 1 with the next ranked documents being closely related linguistically, and often topically. The procedure then seeks to correct mistranscriptions in the topmost ranked document using words within a preset edit distance contained in the next R_1 documents. The string-edit distance measure is used to compare each word in the top-ranked document to all words satisfying preset criteria in the documents ranked below them. These criteria are as follows. A candidate word must appear $\geq m$ times in the $R_1 - 1$ documents below rank 1 (since the item at rank 1 is the query itself), where $m = \sum_{k=2}^{R_1} tf(i,k)$ where k represents the documents containing the candidate combination terms. We also impose the constraint that only terms with identical first letter are allowed candidates; failure to do this was found to introduce too many incorrect candidates. Words satisfying these constraints and within an edit distance e are then added to the query document. The assumption being that if they are sufficiently frequent in the context of related documents and look similar to the term under consideration, then they are probably correct. We also explored the use of fixed values of $n(i)$ as the value of m, but found this to be not sufficiently discriminatory. Incorporating the $tf(i,j)$ rather than just binary presence/absence in m means that we capture multiple occurrences of a candidate word string closely related to the potential mistranscription, even if it only occurs in a very small number of documents matching the document query. The following is a short example snippet of a document with the inserted "corrections" shown in bold,

" ... look at out top stories - dosabled **disabled** gopfer **golfer** casey **case** martin won right drive cart **case** professional tnur. **tour** pga argued cart **case** gives martin unfair advamtage ..."

It can be seen that a number of accurate corrections are made, although some errors are made for short words, and no insertion is made for the term "advamtage" since candidates did not appear in the closely matching documents. "Advantage" is unlikely to be a topically specific term in this context and its appearance in related documents is thus likely to be a matter of chance.

This technique is similar to the document expansion technique described in [16] for SDR, but our method focuses on seeking to correct errors in individual elements of the identified content of the documents based on their character structure rather than using overall collection level statistics to select terms that are likely to have occurred in a document.

Tables 3, 4 and 5 show results for index-time combination with several settings of m and R_1, where $e < 4$ in all cases. These values were chosen after an extensive set of experiments with a subset of the test collection. Interestingly the value of $e < 4$ is the same as that which generally works best for the query-time technique. The tables show new baseline results which are needed since the features of the document collection have been changed. While the new AvP baseline figures here are only marginally higher than those in the original baseline in Table 1, results for 5 and 20 expansion terms show improvement in all retrieval measures. The change for the PRF runs is shown relative to the new baseline in each case. The method produces a marginal improvement for TR relative to Table 1, but PRF is now effective for DIR, although the absolute results are slightly lower than those in Table 2 using the query-time method. Using the indexing-time correction method there is now an improvement in retrieval performance for SDR

Table 6. Index-time additions to 50 sample documents for Text, Speech and OCR collections.

	Text	Speech	OCR
Identified Potential Errors	159	168	318
Correct Additions	54	57	95
False Positives	105	111	219

compared to that in Table 1. While it may appear obvious that correction of the index file will improve retrieval effectiveness, the degree of change is not easily predictable. As we see here, while it only produces a small change in baseline retrieval performance, its effect on PRF is much more dramatic.

In order to explain these results more fully, we analyzed the behaviour of the correction method. We randomly selected 50 news story documents and extracted these for each of the document sets used to generate the results in Table 5. For each document we assessed each identified correction in the combined transcriptions. Results of this analysis are shown in Table 6 from which it can be seen that the number of corrections average around one per document for the Text and Speech data and two per document for the OCR data. A high number of False Positives appear in all cases. However, many of these are words strongly related to the correct word (e.g. "Yugoslav" appearing in place of "Yugoslavia", and similarly "Buddhism" for "Buddhist") which when stemmed will function as the correct search term. We also noted that on a number of occasions a word is added which, while not present in the original document, proves to be very useful for retrieval (e.g. "rifle" being combined with "right"). This last result indicates that there may be benefit in exploring document expansion methods further [16]. The number of corrections for the Text documents is perhaps unexpectedly high. However, it should be remembered that these transcriptions contain spelling mistakes and actual manual errors in transcriptions, as noted in Section 4.1.

A number of the false positives are short words unrelated to the contents of the document, and their presence in the document index may damage retrieval effectiveness. In order to reduce the number of false positives we imposed a further constraint on the index used to filter out combination words with < 6 characters. The length constraint was found to significantly reduce the number of false positives, but also the number correct additions. In retrieval experiments it was generally found to be more effective not to apply this length constraint, lack of space prevents us from reporting these results here.

Combining Index-Time and Query-Time Term Combination Table 7 shows results of using the indexing-time combined collection from Table 5 with 20 expansion term PRF using query-time merging. The results follow similar trends with respect to the maximum edit distance to those in Table 2. In all cases any improvements over the results in Table 5 are very small, and absolute values are no better than those achieved for query-time only combination in Table 2. However, the improved result for SDR in Table 5 is preserved after the query-time combination. Overall though using both methods in combination is probably not justified computationally given the small variations from the results for the methods in isolation.

Table 7. Results using indexing-time term combination ($e < 4$, $m > 4$, $R_1 = 10$) with 20 expansion and query-time string-comparison term merging.

Media	TR				SDR				DIR			
MaxEd	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet	P10	P30	AvP	RelRet
1	0.600	0.409	0.522	1538	0.563	0.391	0.483	1541	0.563	0.388	0.502	1598
chg bl. (%)	+11.3	+10.8	+10.6	17	+14.0	+16.7	+17.2	+54	+4.5	+8.4	+8.2	+46
2	0.598	0.409	0.523	1582	0.565	0.391	0.487	1549	0.576	0.390	0.499	1563
chg.bl. (%)	+10.9	+10.8	+10.8	+27	+14.4	+16.7	+18.2	+62	+6.9	+8.9	+7.5	+11
3	0.606	0.401	0.520	1604	0.561	0.388	0.482	1518	0.582	0.389	0.503	1580
chg. bl. (%)	+12.4	+8.7	+10.2	+49	+13.6	+15.8	+17.0	+31	+8.0	+8.7	+8.4	+28
4	0.586	0.398	0.507	1592	0.559	0.391	0.484	1548	0.569	0.391	0.508	1568
chg. bl. (%)	+8.7	+7.9	+7.4	+37	+13.2	+16.7	+17.5	+61	+5.6	+8.4	+9.5	+16
5	0.582	0.394	0.511	1575	0.559	0.393	0.488	1529	0.586	0.391	0.501	1590
chg. bl. (%)	+8.0	+6.8	+8.3	+20	+13.2	+17.3	+18.4	+42	+8.7	+8.4	+8.0	+38

6 Conclusions and Further Work

Query-time and index-time methods have been described and evaluated using string-comparison in context to improve PRF for text retrieval, SDR and DIR. Positive results have been demonstrated on a parallel collection of text, speech and paper documents based on the TREC-8 SDR task. We are currently exploring the use of our document correction method for the CLEF speech retrieval task based on oral testimonies [6]. Following the promising results for text retrieval in this paper, we also intend to explore the application of these query and document combination techniques for term correction on larger text retrieval tasks. Preliminary results using the query-time method with the TREC-7 ad hoc search task indicate that it gives an improvement over results achieved using our standard information retrieval with PRF system. We believe that the results and methods described here easily extend to true relevance feedback, and we aim to demonstrate this in further work.

While our results so far are very encouraging, we can expect them to improve further if the correction methods are made more reliable. At present these make no formal use of linguistic context or data from the recognition process. A possible means to improve the correction methods accuracy could be to make use of statistical language models to give a quantitative measure of the likelihood of a potential correction term appearing in a particular place within a document, and the recognition likelihood data from speech recognition or OCR, or a statistical estimate of likely character string substitutions in combination with string-edit distance measures. Interesting methods using content correction techniques of this type have previously been reported in [17] [18].

Finally, we plan to explore the application of the results of this study for alternative information retrieval approaches such as query expansion when using language modelling methods.

References

[1] A. M. Lam-Adesina and G. J. F. Jones. Examining and Improving the Effectiveness of Relevance Feedback for Retrieval of Scanned Text Documents. *Information Processing and Management*, 43(3):633–649, 2006.

[2] `trec.nist.gov`

[3] `ir.nist.gov/ria/`

[4] J. S. Garafolo, C. G. P. Auzanne and E. M. Voorhees. The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access*, pages 1–20, Paris, 2000.

[5] S. E. Johnson, P. Jourlin, K. Sparck Jones and P. C. Woodland. Spoken Document Retrieval for TREC-8 at Cambridge University. In *Proceedings of the Eighth Text REtrieval Conference (TREC-9)*, pages 157–168, Gaithersburg, MD, 2000. NIST.

[6] R. W. White, D. W. Oard, G. J. F. Jones, D. Soergel and X. Huang. Overview of the CLEF-2005 Cross-Language Speech Retrieval Track. In *Proceedings of the CLEF 2005 Workshop*, Vienna, 2005.

[7] P. B. Kantor and E. M. Voorhees. The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval*, 2:165–176, 2000.

[8] K. Taghva, J. Borsack, and A. Condit. Evaluation of Model-Based Retrieval Effectiveness with OCR Text. *ACM Transactions on Information Systems*, 14(1):64–93, 1996.

[9] G. J. F. Jones and A. M. Lam-Adesina. An Investigation of Mixed-Media Information Retrieval. In *Proceedings of the 6th European Conference on Research and Development for Digital Libraries*, Rome, pages 463–478, 2002, Springer.

[10] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu and M. Gatford. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST, 1995.

[11] A. M. Lam-Adesina and G. J. F. Jones. Applying Summarization Techniques for Term Selection in Relevance Feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–9, New Orleans, 2001. ACM.

[12] C. Auzanne, J. S. Garafolo, J. G. Fiscus and W. M. Fisher. Automatic Language Model Adaptation for Spoken Document Retrieval. In *Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access*, pages 1–20, Paris, 2000.

[13] G. J. F. Jones and M. Han. Information Retrieval from Mixed-Media Collections: Report on Design and Indexing of a Scanned Document Collection. Technical Report 400, Department of Computer Science, University of Exeter, January 2001.

[14] E. Mittendorf and P. Schauble. Information Retrieval can Cope with Many Errors. *Information Retrieval*, 3:189–216, 2000.

[15] J. Zobel and P. Dart. Phonetic String Mathing: Lessons from Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, pages 30–38, 1996, ACM.

[16] A. Singhal and F. C. N. Pereira. Document Expansion for Speech Retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, pages 34–41, 1999, ACM.

[17] X. Tong and D. Evans. A Statistical Approach to Automatic OCR Error Correction in Context. In *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen, pages 88–100, 1996.

[18] K. Collins-Thompson, C. Schweizer and S. Dumais. Improved String Matching Under Noisy Channel Conditions. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, Atlanta, pages 357–364, 2001, ACM.