Search of Spoken Documents Retrieves Well Recognized Transcripts

Mark Sanderson, Xiao Mang Shou

Department of Information Studies, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK
{m.sanderson, x.m.shou}@shef.ac.uk

Abstract. This paper presents a series of analyses and experiments on spoken document retrieval systems: search engines that retrieve transcripts produced by speech recognizers. Results show that transcripts that match queries well tend to be recognized more accurately than transcripts that match a query less well. This result was described in past literature, however, no study or explanation of the effect has been provided until now. This paper provides such an analysis showing a relationship between word error rate and query length. The paper expands on past research by increasing the number of recognitions systems that are tested as well as showing the effect in an operational speech retrieval system. Potential future lines of enquiry are also described.

1. Introduction

The Spoken Document Retrieval (SDR) track was part of TREC from 1997 (TREC 6) to 2000 (TREC 9). During this period, substantial research and experimentation was conducted in speech retrieval. The work focused on retrieval of radio and TV broadcast news: high quality recordings of generally clearly spoken scripted speech. The overall result of the track (as reported in the summary paper by Garofolo et al, 2000) was that retrieval of transcripts generated by a speech recognition system was almost as effective as retrieval of transcripts generated by hand with proper expansion techniques. Garofolo et al also presented results showing that there appeared to be a relationship between WER and retrieval effectiveness. They showed that for topics where retrieval was effective, WER of retrieved items tended to be low. The authors speculated that hard to recognize documents may also be hard to retrieve.

A more detailed analysis of the reasons for the success of spoken document retrieval was described by Allan in his review of SDR research (2002). Allan pointed out that documents that were most relevant to a query were ones that had query words repeated many times (i.e. the words had a high term frequency - *tf* - within the document). The repetition of query words within a document provided to the recognition system multiple opportunities to spot the query words correctly. Documents that contained query words only once may not have had such word occurrences spotted by a recognizer and therefore were less likely to be retrieved, however, such documents were also less likely to be relevant to the query; failing to retrieve them was not particularly important. Actually the failure to recognize single occurrences of terms in non relevant documents may offer an advantage in SDR over text retrieval as the speech document will not be retrieved. Allan reported that retrieval from spoken document collections with a high Word Error Rate (WER) resulted in poorer effectiveness than that resulting from retrieval over a collection with a low WER. Allan also reported that this inverse relationship between WER and retrieval effectiveness was linear.

Following on from those the two review papers, additional analysis of the SDR track data was conducted by Shou, Sanderson and Tuffs (2003) who reported work describing the variation of the word error rates of retrieved documents across ranking. In the paper, it was shown that across the groups who submitted runs to the TREC SDR track, top ranked documents in each run had a lower WER than documents that were further down the ranking. Some speculations on the reasons for this effect were provided, but little evidence of a reason was reported. This paper provides such evidence.

The paper starts with an overview of past work, followed by a series of experiments that expand on the work reported in the 2003 paper.

2. Past work

Beyond Garofolo et al’s observation of a relationship between effective topics and WER, little past work on the relationship between effectiveness, document rank and word error in recognized transcripts has been reported. However, some related research has been published, which is now described.

In the internal working of a speech recognition system, an audio segment of speech is recognized into a lattice of possible text strings, each string a hypothesis of what was spoken. The hypotheses are compared to the acoustic and language models stored in the speech recognizer. Based on both models, a confidence score is assigned to each word in each hypothesis, signifying the probability that the word was spoken. The sequence of words with the highest scores is chosen as the text string the recognizer will output. It can be expected that the higher score assigned to a word, the more confident one can be that the recognizer’s selection was correct. Zechner and Waibel investigated summarization of spoken documents (2000) and use of confidence scores to improve summarization quality. Their summarizer ranked passages of a spoken document by their similarity to the overall document. Summary quality was computed by counting the number of relevant words (manually identified in human transcription) found within the summary. It was found that if the ranking formula was adjusted to prefer passages holding words with high confidence scores, the quality of the summaries increased by up to 15%. With Zechner and Waibel an approximation of word error rate (i.e. the confidence scores) was used to influence a ranking algorithm to improve the quality of the top ranked passages. Given such success, one might assume that similar use of confidence scores in information retrieval ranking algorithms would also be beneficial. However, attempts to improve retrieval effectiveness through use of the scores have at best been marginally successful (see Siegler et al, 1998, Johnson et al, 1999).

Sanderson and Crestani conducted preliminary investigations of retrieval from a collection composed of both hand transcribed (containing only human errors) and speech recognized documents (with a level of word error within them) (1998). Two versions of each spoken document were placed into a collection, one hand transcribed and one speech recognized. By having pairs of identical documents in the collection, the only difference in the two sub-sets of the mixed collection was the errors in the speech recognized set. If one was to retrieve on such a collection, any difference in rank positions of documents from the two subsets would be due to the error in the second set. Sanderson and Crestani reported that retrieval from such a collection resulted in the hand transcribed documents being retrieved at higher rank positions than the speech recognized documents. By experimenting with two retrieval ranking algorithms, Sanderson and Crestani were able to show the predominant reason for the hand transcribed documents being ranked higher than the recognized was due to word errors reducing the *tf* weight assigned to words in the recognized documents, therefore making such a document receive a lower score than that assigned to hand transcribed when ranking documents relative to a query. Sanderson and Crestani assumed that documents in the recognized collection had a uniform word error rate and did not explore the effect of different word error rates across such a collection. Neither was the investigation run across a range of retrieval systems or outputs from other speech recognition systems. Further research in retrieval from similar forms of collection was conducted by Jones and Lam-Adesina (2002).

3. Experiments on the extent of the effect of WER and rank position

In their paper, Shou, Sanderson and Tuffs (2003) presented evidence of variation of WER across rankings. That work is expanded on here. In the past paper, the speech recognized transcripts of the one hundred hours of audio data making up the TREC-7 SDR collection were collected from six of the groups participating in the speech track. In addition, the runs submitted by each group were also gathered: these hold the ranked list of documents retrieved for each topic by each group’s retrieval system. The collection had an accompanying accurate manually generated text transcript, which allowed WERs to be computed for each document at each rank position for each topic within each collected transcript. A scatter plot of the WER of retrieved documents against their rank position was produced for each of the six transcripts. In addition, one of the six transcripts, from AT&T, had two forms of retrieval system search over it, which resulted in seven plots. The seven data sets are now described.

1. derasru-s1, UK Defence Evaluation and Research Agency (DERA, Nowell, 1998). Here a large vocabulary continuous speech recognizer (50,000 word vocabulary plus 500 bigrams) developed by DERA was used to generate the transcript. Its average word error rate was 66.4%. Retrieval was based on the Okapi system. The topics of the TREC track were syntactically tagged. Certain syntactic patterns were used to identify keywords of the topic text. Selected topic keywords were expanded with synonyms and sometimes with hypernyms taken from the WordNet

thesaurus. When keywords were ambiguous, the commonest synset was chosen to provide expansion terms.

2. derasru-s2: using the same retrieval set up as derasru-s1, the speech recognizer had an additional processing step, which reduced the error rate to 61.5%. Here the audio data was segmented into different streams depending on the quality of audio recording found within parts the TREC spoken document collection. Audio recordings identified as being speech over telephones for example were recognized differently from segments judged to be recorded to a higher quality.
3. att-s1, AT&T. Recognition was performed using an in-house speech recognition system that produced transcripts with a 32.4% WER. The vocabulary size of the system was not stated in the paper describing the AT&T submission to TREC (Singhal et al, 1998). Retrieval was based on the SMART retrieval system with a phrase identification process operating on TREC topic text and pseudo-relevance feedback used to expand topics with additional terms. The form of feedback used was a method referred to as collection enrichment: here the first search of the pseudo-relevance feedback stage was conducted on a large collection of news articles and not the relatively small SDR collection.
4. att-s2. For the second AT&T submitted run, the same recognition system was used, but retrieval was altered to include a document expansion step. Here in the same manner that topic text was expanded using pseudo relevance feedback, each recognized transcript was expanded, by searching a large collection of newspaper texts with the transcript text as a query. The transcript was expanded with terms found to commonly co-occur in top retrieved newspaper articles. This run produced better retrieval results than att-s1.
5. dragon-s1, Dragon systems and the University of Massachusetts. This was a combined submission using a speech recognizer from Dragon and retrieval using the UMass Inquery retrieval system (Allan et al, 1998). The recognizer used a 57,000 word vocabulary. It produced transcripts with an error rate of 29.8%. Prior to retrieval, topic text was processed to locate phrases, which were then searched as phrases. Certain proper nouns were expanded with synonyms. A form of pseudo relevance feedback (known as local context analysis) was used to expand topic texts with additional terms taken from the recognized transcript collection.
6. shef-s1, University of Sheffield with collaborators at Cambridge University (Abberley et al, 1998). Recognition was performed using the Abbot recognizer system with a vocabulary of 65,532 words producing a transcript with a 35.9% WER. Retrieval was performed with a locally built IR system using Okapi-style BM25 weights.
7. cuhkt-s1, University of Cambridge (Johnson et al, 1998). Recognition was performed using the HTK speech toolkit recognizing from 65,000 word vocabulary. The resulting transcript had a 24.8% WER. Retrieval used the Okapi system using BM25 weights. Expansion of selected topic terms with synonyms and with additional terms using pseudo-relevance feedback was used, as was phrase spotting in topic text. Matches on proper nouns and nouns were preferred over adjectives, adverbs and verbs as this strategy was found to bring improvements in retrieval effectiveness.

As can be seen from the descriptions, the seven runs represent a relatively diverse set of retrieval and recognition approaches. The average WER of the transcripts ranged from 24.8% to 66%. Note that two further recognizer transcripts were produced and archived in this year of TREC, nist-b1 and nist-b2 (Garofolo et al, 1999). However, no associated retrieval runs performed on these transcripts were located and so were not used in this experiment.

3.1 The experiment

For each run, rankings for each of the 23 topics (51-73) were gathered from the TREC web site. NIST's sclite software was used to calculate the WER of each document retrieved in the top 200 rank positions. Since sclite only calculates WER based on speaker id, the original recognized transcripts were modified by replacing speaker ids with document ids so that WER could be measured on each document. After obtaining WER of each story across all systems, the average error at each rank position across the 23 queries was calculated and graphed.

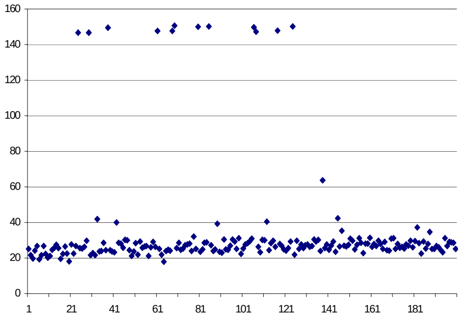

Figure 1: document rank (x-axis) vs. word error rate (y-axis) for dragon-s1 system

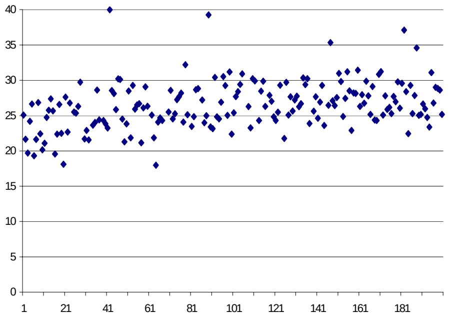

Figure 2: Graph of Figure 1 with y-axis adjusted to focus on majority of retrieved documents.

The graph (in Figure 1) shows a slight increase in error rate for recognized documents at higher ranks. A small set of documents with a very a high error rate across the ranking was observed (the twelve points at the top of the scatter plot). The reason for this effect was investigated and found to be related to mistaken insertions of large amounts of text into short documents by the recognizer (such erroneous documents were found in all six transcripts). Ignoring these few high error rate documents by focusing the scatter plot on the main band of documents reveals the trend of increasing error rate more clearly. It can be seen that top ranked documents (those on the left side of the graph) have a lower word error rate than those ranked further down the ranking. The plot such as that shown in Figure 2 was repeated for all other six runs and is displayed in Figure 3 – 8. Across all runs, the average WER for the very top ranked documents (those in the top 10) is lower than the WER for documents in the wider part of the ranking. Such differences in WER are also shown in Table 1 where the average WER is calculated in the top 10, 50 and 200 rank positions and it can be seen that for all recognizers and runs WER is lower for higher ranked documents.

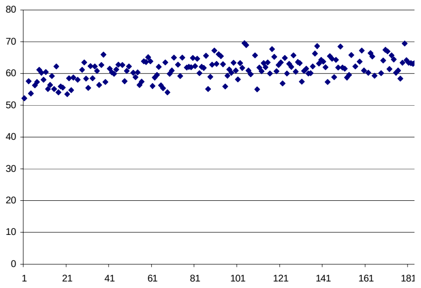

Figure 3: derasru-s1, rank vs. WER

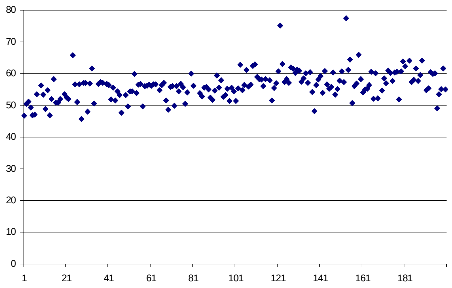

Figure 4: derasru-s2, rank vs. WER

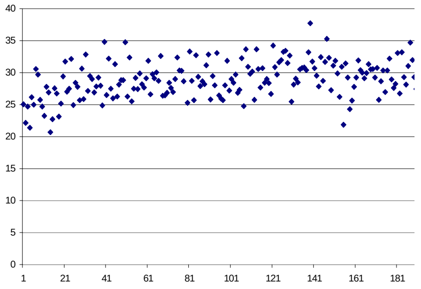

Figure 5: att-s1, rank vs. WER

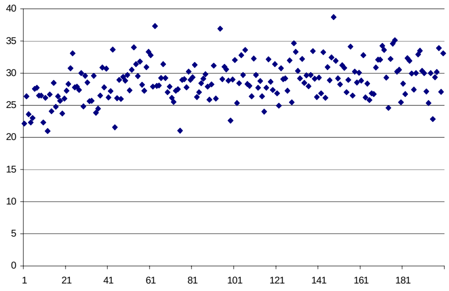

Figure 6: att-s2, rank vs. WER

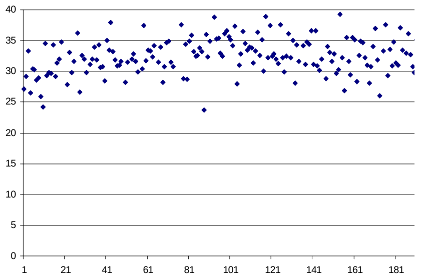

Figure 7: shef-s1, rank vs. WER

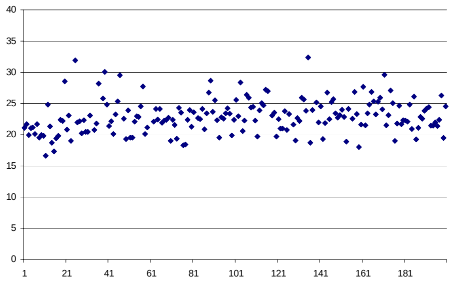

Figure 8: cuhtk-s1, rank vs. WER

Run	Average WER in top 10 (%)	Average WER in top 50 (%)	Average WER in top 200 (%)
derasru-s1	57.1	58.8	62.8
derasru-s2	50.5	53.1	56.2
att-s1	25.5	27.4	29.1
att-s2	24.8	26.8	29.0
dragon-s1	22.8	25.1	26.9
shef-s1	28.4	32.7	33.4
cuhtk-s1	20.6	22.2	23.4

Table 1. Word Error Rate differences for top 10, 50 and 200 retrieved documents.

The slight, though consistent trend measured across all data sets provides evidence that when retrieving speech recognized documents, those with lower word error rates tend to be ranked higher. The trend also appears to occur independent of the mix of retrieval strategies used across the runs (e.g. different weighting schemes, use of pseudo-relevance feedback, use of document expansion, etc) and independent of the accuracy of the speech recognizer used.

Although the trend is consistent across the data sets, it is not immediately clear what the cause of such a trend is: one explanation is that top ranked documents tend to contain a broader range of query words than those documents ranked lower. Another explanation mentioned by Sanderson and Shou (2002) is that transcripts of spoken documents containing query words assigned a high *tf* weight – which tend to be ranked highly by retrieval systems – often have a lower overall WER. Determining which of these possible causes might explain the observed effect was the subject of the next experiment.

4. Determining the cause of the effect

As valuable as it can be to examine the search output of other research groups' retrieval systems (as was conducted in Section 3), analyzing the ranked output of a system that one has no access to is often limiting. This is because a common consequence of such analysis is the discovery that new experiments need to be conducted to generate different versions of the data, which requires access to the retrieval system of other research groups, something that is rarely possible. Therefore, in order to conduct more detailed analysis of WER in retrieved documents, the six recognized transcripts used in the experiment of Section 3 along with the two NIST transcripts (nist-b1 and nist-b2) were indexed and searched so that new search output could be created for further experimentation. The aim of the experiments was to examine the relationship between WER, *tf* weights and the number of words in common between a query and a document.

In the experiment, the average WER of top ranked documents retrieved by queries of different length was measured. The TREC-7 SDR collection holds only 23 topics. In order to produce a larger number of topics of different lengths, (non-stop) words were randomly sampled without repeated words from each of the topics. The number of words sampled was varied, producing sets of topics of length 1, 2, 5, 10 and 15. Each of the 23 topics was sampled 1,000 times for each of the five different lengths. The queries were submitted to two versions of the GLASS search engine, an in house IR system that implements Robertson et al's BM25 ranking algorithm (1995) as well as a simple quorum scoring (coordination level matching) algorithm that ranks documents by the number of query words found in a matching document (making no use of *tf*, *idf* weights or of document length normalization). No relevance feedback or other expansion methods were employed in both algorithms. The tables of the results of this experiment are shown in Table 2 and Table 4, which record average WER and Table 3 and Table 5, which display precision measured at rank ten.

Topic length	cuhtk-s1	dragon98-s1	att-s1	shef-s1	nist-b1	nist-b2	derasru-s2	derasru-s1
1	19.3	22.0	22.2	25.8	26.7	39.5	51.4	52.3
2	18.9	22.7	22.1	25.9	26.4	39.1	50.6	52.7
5	16.7	21.0	21.1	23.9	24.3	38.0	44.6	48.6
10	15.9	19.7	20.4	22.1	23.5	36.9	42.9	47.1
15	15.5	19.6	19.9	21.2	23.0	36.7	42.1	46.4

Table 2. The average WER measured across the ten top ranked documents retrieved by quorum scoring for each of the 1,000 topics randomly sampled.

Topic length	cuhtk-s1	dragon98-s1	att-s1	shef-s1	nist-b1	nist-b2	derasru-s2	derasru-s1
1	0.06	0.07	0.06	0.06	0.05	0.05	0.06	0.06
2	0.09	0.08	0.09	0.09	0.08	0.07	0.07	0.07
5	0.20	0.18	0.19	0.19	0.18	0.15	0.17	0.17
10	0.26	0.23	0.26	0.25	0.24	0.20	0.21	0.23
15	0.28	0.24	0.27	0.27	0.25	0.21	0.21	0.23

Table 3. Precision at 10 measured in the retrieved documents shown in Table 2.

Topic length	cuhtk-s1	dragon98-s1	att-s1	shef-s1	nist-b1	nist-b2	derasru-s2	derasru-s1
1	18.1	21.7	22.3	25.3	25.7	39.0	46.9	51.0
2	17.8	22.2	22.4	25.2	25.2	39.2	46.3	50.7
5	17.2	21.1	22.2	24.4	23.9	38.7	44.3	50.6
10	17.2	20.9	21.9	24.1	24.1	38.8	43.0	49.4
15	16.9	20.7	21.9	23.9	24.1	38.4	42.1	48.9

Table 4. The average WER measured across the ten top ranked documents retrieved by BM25 for each of the 1,000 topics randomly sampled.

Topic length	cuhtk-s1	dragon98-s1	att-s1	shef-s1	nist-b1	nist-b2	derasru-s2	derasru-s1
1	0.10	0.09	0.09	0.09	0.08	0.08	0.08	0.08
2	0.16	0.15	0.15	0.15	0.15	0.13	0.13	0.13
5	0.30	0.28	0.28	0.28	0.28	0.24	0.25	0.27
10	0.36	0.35	0.35	0.35	0.35	0.29	0.31	0.33
15	0.37	0.36	0.35	0.36	0.36	0.30	0.33	0.33

Table 5. Precision at 10 measured in the retrieved documents shown in Table 4.

As can be seen, across all eight transcripts for both form of ranking algorithm, as the length of topic increases, the WER measured in the top ranked documents reduces, while precision at 10 increases. This effect is consistent for both forms of ranking algorithm used. From the result with the quorum scoring, it can be concluded that the reduction in WER shown in Table 2 was caused by the change in top ranked documents: as topic length increases the top ranked documents hold more query words. Documents that match on a broader range of query words tend to have a lower WER. While a relationship between the rank position of recognized documents and their WER was observed in the past, to the best of our knowledge a causal effect has not been determined before. From the results in Table 2, we conclude that the process of retrieval itself is locating documents that have a lower WER.

Topic length	av. WER BM25	av. WER Quorum	difference	ttest (p)
1	31.3	32.4	-1.2	0.058
2	31.1	32.3	-1.2	0.057
5	30.3	29.8	0.5	0.099
10	29.9	28.6	1.4	**0.001
15	29.6	28.1	1.6	**0.001

Table 6. Comparison of average word error rate (WER) measured across the eight transcripts shown in Table 2 and in Table 4

The number of words in common between a document and a query is not the full story, however, as it can be seen that for topics of length one, for all but one transcript, WERs are lower using BM25 ranking (Table 4) than when using quorum scoring (Table 2). Here, top ranked documents retrieved by a single word query using BM25 are those documents in the collection that contain the query word repeated the most number of times (normalized by document length). Observing a query word repeated many times in a document would appear to be an indicator that that document was recognized well. The comparison of WERs is summarized in Table 6. The amount of WER reduction is relatively small and for topics of length one or two the difference is not significant. In comparing the error rates across the two ranking algorithms for longer queries (five, ten or fifteen words) the quorum scoring algorithm retrieves documents with lower WERs and for the longest queries lengths, the differences between quorum and BM25 are significant.

However, it must be remembered that quorum scoring though retrieving documents with low WERs is not retrieving the most relevant documents as across the Tables, precision at ten is consistently higher for BM25 ranking. We believe that this effect is due to BM25 top ranked documents matching on fewer query words than the documents top ranked by quorum scoring but with higher *tf*, which means a query word is repeatedly recognized, so BM25 has the effect of ranking higher documents with fewer matching terms.

5. Experiments with manual calculation of WER on top ranked SpeechBot snippets

To provide further confirmation of the results in Section 4, measurements were made of the word error rate in the snippets of top ranked transcripts retrieved by a publicly available spoken document retrieval system, SpeechBot (Van Thong et al, 2000; Moreno et al 2000). We would like to test whether the correlation between word error rate and document ranking could be generally applied to other systems using different speech recognition technologies. A white paper published on the engine’s Web site (Quinn, 2000) described that the engine indexed streaming spoken audio using a speech recognizer. Several thousand hours of audio data were crawled and stored in a searchable collection composed of mainly US-based radio stations producing predominantly news, current affairs and phone-in shows. The snippets in the result list summary presented by SpeechBot were brief sections of speech transcript that strongly matched a user query; most likely selected by a within document passage ranking approach.

The WER of each retrieved snippet was computed by manually comparing the snippet text with human listening to the corresponding part of the audio recording noting any inserted, deleted or substituted words. The WER was calculated using the total number of errors divided by the total number of words in the returned snippets. This method was consistent with NIST’s WER calculation tool sclite which was used in the TREC SDR track. Because the majority of the SpeechBot collection was audio news, 34 current affair queries were created for the experiment. The number of words in the examined snippets ranged from twenty to forty. It was found that audio files were not available with some of the retrieved results (usually occurring with old audio recordings dated before 1999 or with the recordings of certain shows). The authors were made aware that a number of the transcripts used by SpeechBot for certain radio programs like PBS’s News Hours are manually written transcripts and not generated by an SR system (Quinn, 2000), such transcripts were also ignored. Therefore, 311 out of a possible 350 snippets were assessed, the average WER measured within the snippets was 19.29%, and the standard deviation was 14.04%. Among the snippets, the maximum calculated WER was 68.75% while the minimum was 0%. The measured rate was substantially lower than the estimated 50% WER reported to exist across the whole SpeechBot collection (Quinn, 2000). This constitutes further evidence of the retrieval process assigning high rank to well recognized documents.

6. Conclusions and future work

This paper described experiments that demonstrated that when there is variability in the word error rate across the documents of a speech recognized collection, retrieval systems tend to retrieve highest documents with low word error. This effect was demonstrated through experimentation on an operational spoken document retrieval system as well as a series of analyses across multiple speech recognizers and retrieval algorithms. It was shown that documents holding many query words tend to have low WER.

We plan to extend our investigation to other retrieval research areas where documents containing varying levels of error are retrieved. Research topics such as retrieval of transcripts produced by Optical Character Recognition (OCR) of scanned document images or retrieval of documents translated into a different language may be worthy of further investigation. When retrieving OCR’ed documents, understanding if the top ranked are more readable or are the product of a better scan would be a straightforward experiment to undertake. A potentially more intriguing question is if in the context of cross language information retrieval, if top ranked documents are better translated than those retrieved further down the ranked list. To the best of our knowledge, this question has not been addressed within the cross language research community.

References

Abberley, D., Renals, S. and Cook, G. (1998) Retrieval of broadcast news documents with the THISL system; In Proceeding IEEE ICASSP, 3781-3784

Allan, J., Callan, J. Sanderson, M., Xu, J. (1998) INQUERY and TREC-7 in the proceeding of the 7th Text REtrieval Conference (TREC 7)

Allan, J. (2002) Perspectives on Information Retrieval and Speech, in Information Retrieval Techniques for Speech Applications, Coden, Brown and Srinivasan, editors, Lecture Notes in Computer Science, Volume 2273 1-10.

Garofolo, J.S., Voorhees, E.M., Auzanne, C.G.P., Stanford, M., Lund, B.A. (1999) TREC-7 Spoken Document Retrieval Track Overview and Results, in the Proceedings of the DARPA Broadcast News Workshop

Garofolo J.S., Auzanne, C.G.P., Voorhees, E.M. (2000) The TREC Spoken Document Retrieval Track: A Success Story; Proceeding of RIAO

Johnson, S.E., Jourlin, P., Moore, G.L., Spärck Jones, K., Woodland, P.C. (1998) Spoken Document Retrieval For TREC-7 At Cambridge University, in the proceeding of the 7th Text REtrieval Conference (TREC 7)

Johnson, S.E., Jourlin, P., Spärck Jones, K., Woodland, P.C. (1999): Spoken Document Retrieval for TREC-8 at Cambridge University, in the proceedings of the 8th Text REtrieval Conference (TREC 8)

Jones, G.J.F. and Lam-Adesina, A.M. (2002) An Investigation of Mixed-Media Information Retrieval, in the proceedings of the 6th European Conference on Research and Development for Digital Libraries (ECDL), 463-478

Moreno, P., Van Thong, P.M., Logan, B., Fidler, B., Maffey, K., Moores, M. (2000) SpeechBot: A Content-based Search Index for Multimedia on the Web, in the proceedings of the 1st IEEE Pacific-Rim Conference on Multimedia, (IEEE-PCM 2000)

Nowell, P. (1998) Experiments in Spoken Document Retrieval at DERA-SRU, in the proceeding of the 7th Text REtrieval Conference (TREC 7)

Quinn, E. (2000) SpeechBot: The First Internet Site for Content-Based Indexing of Streaming Spoken Audio Technical Whitepaper, Compaq Computer Corporation, Cambridge, Massachusetts, USA

Robertson, S Walker, S Jones, MM Hancock-Beaulieu (1995) Okapi at TREC-3, in the proceeding of the 3rd Text REtrieval Conference (TREC 3)

Sanderson, M., Crestani, F, (1998) Mixing and Merging for Spoken Document Retrieval, in the proceedings of the 2nd European Conference on Digital Libraries; Heraklion, Greece. Lecture Notes in Computer Science N. 1513, Springer Verlag, 397-407

Sanderson, M., Shou, X.M. (2002) Speech and Hand Transcribed Retrieval; Lecture Notes in Computer Science N.2273, Information Retrieval techniques for Speech Application, Springer, 78-85

Shou, X.M., Sanderson, M., Tuffs, N. (2003) The Relationship of Word Error Rate to Document Ranking, in the proceedings of the AAAI Spring Symposium Intelligent Multimedia Knowledge Management Workshop, Technical Report SS-03-08, ISBN 1-57735-190-8, 28-33.

Siegler, M., Berger, A., Witbrock, M., Hauptmann, A. (1998): Experiments in Spoken Document Retrieval at CMU, In the proceedings of the 7th TREC conference (TREC-7)

Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F. (1998) AT&T at TREC-7, in the proceeding of the 7th Text REtrieval Conference (TREC 7)

Van Thong, J.M., Goddeau, D., Litvinova, A., Logan, B., Moreno, P. Swain, M. (2000) SpeechBot: A Speech Recognition based Audio Indexing System for the Web in the proceedings of the International Conference on Computer-Assisted Information Retrieval, Recherche d'Informations Assistee par Ordinateur (RIAO2000) 106-115

Zechner, K., Waibel, A. (2000) Minimizing Word error rate in Textual Summaries of Spoken Language, in the proceedings of NAACL-ANLP-2000, 186-193