The Relationship of Word Error Rate to Document Ranking

Xiao Mang Shou, Mark Sanderson, Nick Tuffs

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

Abstract

This paper describes two experiments that examine the relationship of Word Error Rate (WER) of retrieved spoken documents returned by a spoken document retrieval system. Previous work has demonstrated that recognition errors do not significantly affect retrieval effectiveness but whether they will adversely affect relevance judgement remains unclear. A user-based experiment measuring ability to judge relevance from the recognised text presented in a retrieved result list was conducted. The results indicated that users were capable of judging relevance accurately despite transcription errors. This lead an examination of the relationship of WER in retrieved audio documents to their rank position when retrieved for a particular query. Here it was shown that WER was somewhat lower for top ranked documents than it was for documents retrieved further down the ranking, thereby indicating a possible explanation for the success of the user experiment.

Introduction

The Spoken Document Retrieval (SDR) track has been included in TREC from 1997 (TREC 6) to 2000 (TREC 9). During this period there were substantial amounts of research and experiments carried out and some important conclusions were drawn. In general, retrieval of the transcripts of spoken documents was viewed as a success and since TREC 9, SDR became a somewhat "solved problem" and that research efforts should be focused elsewhere [1].

However, following the trend of multimedia retrieval, more and more ASR (Automatic Speech Recognition) Systems are commercially available and audio search has become more and more frequent in information retrieval, application wise, new problems arise which have not yet been covered in previous SDR research, especially in the issues related to the usability of SDR systems. One of the important conclusions throughout the SDR research in TREC is that WER has only a modest impact upon retrieval performance, this is because the length of the transcripts provides enough keyword repetition and enough related contextual material to compensate for the recognition errors ([1], [3], [5], [12]). At present, many speech search engines present a text-based summary or surrogates of retrieved audio documents. In a typical situation like this, users need to assess the surrogates to decide whether a document is relevant or not before they choose to listen to the audio document. Though WER does not seriously affect retrieval accuracy, with a WER over 20% (minimum mean WER across all SDR track collections [1], [3], [5]), a text transcript can be difficult to read and therefore, difficult for users to decide its relevancy. The impact of WER on user relevance judgement using speech surrogates remains unclear and it is our intention to explore in this area. This paper describes such an exploration by describing two experiments each proceeded with a brief overview of pertinent past work.

Experiments on User Relevance Judgement Using Speech Surrogates

Various past work ([1], [2], [3], [4], [5], [6], [7], [8], [9]) has demonstrated that WER in SDR has a limited, influence in retrieval performance. This is due to the fact that the important terms in a document usually repeat more times so if some of them are mistakenly recognised, the rest of the correctly recognised ones can still be used to locate the document. However, though retrievable, the misrecognised words in documents can cause difficulties in reading and understanding, making it hard for users to make relevance judgements using speech surrogates being presented to them by search engines. If this hypothesis can be proved, then some other means of surrogates need to be sought rather than using passages taken from speech transcripts. It is our motivation to investigate the accuracy and usefulness of speech surrogates and observe any impact of WER on influencing user relevance judgements that our experiments were designed. The operational SDR search engine SpeechBot (http://speechbot.research.compaq.com/, [10], [11], [16]) was chosen for our experiments. It is an engine for indexing streaming spoken audio on the Web using an automatic speech recogniser. Several thousand hours of audio data crawled from a variety of mainly US Web sites

is stored in the collection covering news and current affairs.

The surrogates in SpeechBot are brief sections of speech transcripts around matched query words; most likely selected by a within document passage ranking approach (such a successful means of generating a document summary/surrogate was implemented and tested by Tombros and Sanderson [15]). Users need to judge relevance according to these surrogates and if interested, they can listen to the original audio starting from the vicinity of the matched best passage. This is similar to browsing full documents in text search engines.

Our experiments [14] had ten users (University students who were paid a small fee, £15, for participation) use SpeechBot to complete one query (on a current affairs question) and judge the twenty top ranked documents for relevance. Two major facets were applied in comparison: user's relevance judgements and the time used to complete the tasks. User's relevance judgements were measured by the rate of judgement change. In the experiments, users were first asked to use the surrogate information to make a decision about relevance and then the alterations to their original decisions were examined after they listened to the audio files. The rate of judgement change captured the change in percentage of relevant and non-relevant documents and the accuracy of judgement records the unchanged relevant and non-relevant documents in percentage after listening to the audio file. The former measure related to users changing their mind about relevance therefore yielding some clue about how effective and accurate relevance judgements in speech surrogates was compared with listening to the full audio. Tables 1 and 2 list the result of judgement change and judgement accuracy obtained from our experiments. Perhaps surprisingly, the two measures show no large difference between relevant and non-relevant judgements using the text summaries and listening to the audio documents returned by SpeechBot.

Surrogates only		**Full document**	
Relevant	**Non-relevant**	**Relevant**	**Non-relevant**
35.5%	64.5%	43%	57%

Table 1. Change of relevance judgement (SpeechBot)

Relevant	**Non-relevant**
62.8%	85.1%

Table 2. Judgement accuracy after listening to audio files (SpeechBot)

In a post test questionnaire, 50% of users indicated that word errors in the transcripts caused them difficulty in reading and thereby judging relevance, but it would appear that the difficulty did not affect judging accuracy significantly.

The time users took to judge surrogates on SpeechBot was timed also. On average, 10.7 seconds were needed to make a judgment of relevance. In a parallel test, the same users were also asked to examine the surrogates presented in a Google search result list (for a different query). Users were found to take a similar amount of time to judge surrogates: 10.3 seconds per judgement. The result lends some support to the notion that transcription errors were not affecting judgements on SpeechBot result lists.

It was also recorded that on average 36 seconds were used to make a decision on relevance by listening to audio files. Given that users cannot easily scan over an audio document as they might with a text document, it is striking that so little time is required to make a judgement on relevance. However, that they take three times as long to make the judgement compared to surrogates indicates that ensuring the surrogate is as accurate as possible will save significant time for the user.

In a general interview after the questionnaire, users made a few additional pertinent comments about the SpeechBot system in general:

- some complained about different volume levels of the retrieved audio documents requiring continual adjustment of volume as each document was examined;
- the point at which the audio documents were played from was also criticised with some users feeling the starting point was too early;
- one user praised the system highly stating that as she was a dyslexic, listening to relevant documents was much easier than reading similar textual ones.

The conclusion of this experiment was that users were apparently able to make accurate relevance judgements based on the surrogates provided by SpeechBot: their judgement on the surrogates was similar to that made after listening to the original audio source. While admittedly a small sample on which to draw conclusions, the study appeared to produce somewhat unexpected results, given the relatively high levels of word error rate within the documents recognised by SpeechBot. In a white paper describing the system [16], a WER of 50% was estimated to be occurring on average across the collection. Such a level of error should render transcripts

hard or very hard to read, however, the effects of the error were not apparent.

In considering reasons for the success of users in judging relevance, the following idea was considered. If a piece of speech transcript is retrieved in top rank positions, it usually means that query terms are better recognised in this piece and the context surrounding query words, which becomes the surrogate, stands a higher chance to be better recognised as well. Therefore, the top twenty ranked documents used in our experiments could be composed of the best recognised (low WER) documents which are relative easy to read in relevance judgement. Whether and how the WER may increase with rank positions and its impact on the usefulness of surrogates was therefore examined. This was the motivation of our following experiments.

Experiments on the Correlation between WER and Rank Position

Given the previous usability experiment, an ideal follow-on for work of this kind would be to determine the word error rate in documents retrieved by SpeechBot for a set of queries. The idea of this experiment was based on the assumption that the top ranked documents were also some of the better recognised documents. Our experiment was designed to calculate the average WER within each retrieved document examining if the rate correlated with rank position. To achieve this with SpeechBot, hand-transcripts of spoken documents and an automatic tool to calculate WER between speech documents and hand transcripts on each document basis were required. However, such an experiment would require obtaining or generating a large number of written transcripts from the retrieved speech documents in order to determine a document by document WER. As this would involve a great deal of effort, as a first step, a preliminary experiment to test the idea that was relatively easy to execute was required.

Here the one hundred hour TREC-7 SDR collection proved to be of use as it holds a high quality manually generated text transcript already created, as well as seven speech recognised transcripts each retrieved using different search systems with the rankings from these searches recorded on the TREC web site. The seven recogniser/retrieval system pairs used were derasru-s1, derasru-s2 (Defence Evaluation and Research Agency, UK [8]), att-s1, att-s2 (AT&T [7]), dragon-s1 (University of Massachusetts retrieval system & Dragon systems recogniser [9]), shef-s1 (Sheffield University using Abbot system [2]), and cuhkt-s1 (Cambridge University using HTK toolkit [6]). Rankings for query topics 51 to 73 were gathered for the systems from the TREC web site. NIST's sclite software (http://www.nist.gov/speech/tools/) was used for calculating WER. Since sclite only calculates WER based on speaker id, the original recognised transcripts had to be modified by replacing speaker ids to story ids so that WER could be measured on each story (document) basis. After obtaining WER of each story across all systems, the average of them in each rank position across the 23 queries could then be calculated.

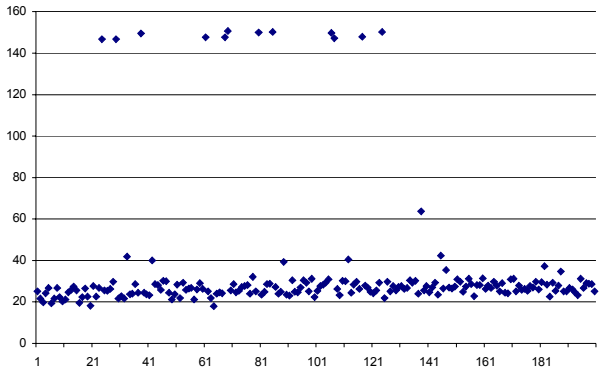

Figure 1: document rank (x-axis) vs. word error rate (y-axis) for dragon-s1 system

The graph in figure 1 appears to show a slight increase in error rate for recognised documents at higher ranks. There also appears to be a small set of documents that have a high though consistent error rate across the ranking (the twelve points at the top of the scatter plot). The exact reason for this effect appears to be related to mistaken insertions of large amounts of text into short documents, however, the effect is not fully understood and is the subject of further work. Ignoring these few high error rate documents by focussing the scatter plot on the main band of documents reveals the trend of increasing error rate is present across all systems with the possible exception of the Cambridge system, as shown in Figure 2.

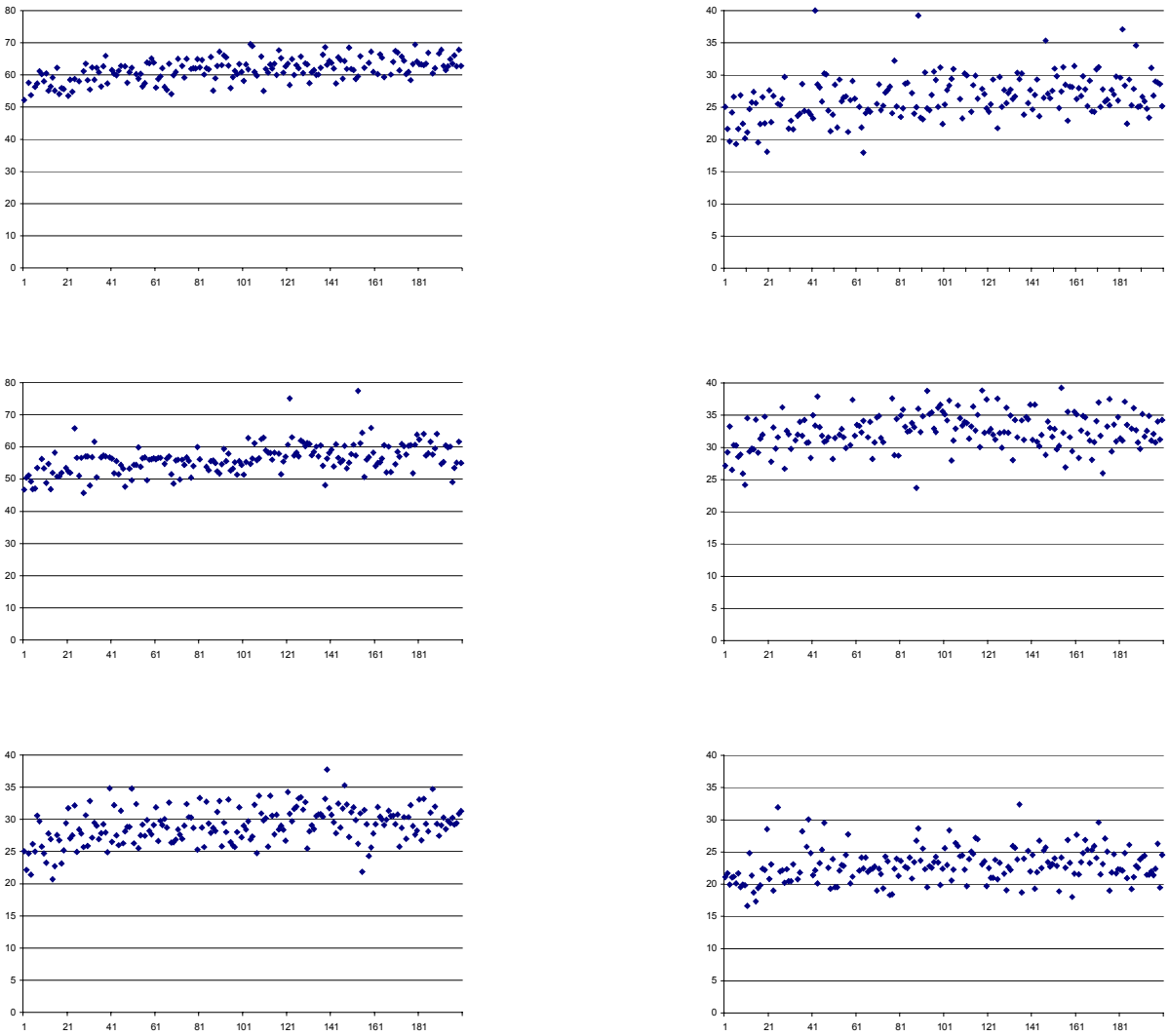

Figure 2: Scatter plots of rank vs. word error rate for TREC-7 SDR runs derasru-s1, derasru-s2, att-s1, att-s2, dragon-s1, shef-s1, and cuhtk-s1

Such a consistent, though slight, trend across all systems gives us some confidence in stating that when retrieving speech recognised documents, those with lower word error rates tend to be ranked higher. This, in turn should give one confidence in the readability of transcripts of top ranked documents even if the average word error rate across a collection is low. Such a result complements the usability experiment described above. However, though there is a increase in WER in top rank positions, it is only slight, however, the data used in our WER experiments are TREC-7 speech transcripts recognised from broadcast news and not the apparently noisier data that SpeechBot uses. We anticipate that with a larger, more heterogeneous, and noisier data set, the error rate rank trend maybe more pronounced. To investigate this effect in SpeechBot rankings is part of our planned future work.

Conclusion and Future Work

This paper described experiments carried out to measure the impact of WER in relevance judgements using text-based speech surrogates. The results showed that speech surrogates alone were sufficient to make effective relevance judgements and recognition errors did not have significant adverse impact on the judgements. In the second experiment, a slight trend was detected in variation in WER against rank position: the higher the rank, the higher the word error rate.

As already stated, we plan to examine other data sets, such as SpeechBot's collection, to explore the WER/rank relationship further. Given that the focus of our interest has been on the best passages of spoken documents presented in retrieval result lists, we are also interested to examine error rates in the best passage as opposed to the document as a whole. Such a finer grained analysis may reveal a stronger relationship than currently observed. It may also be worth examining other retrieval research areas where documents with varying levels of error in them are retrieved. Research topics such as retrieval of scanned document images or retrieval of documents translated into a different language may be worthy of further investigation.

Acknowledgements

A portion of this work was conducted as part of one of the author's Masters dissertation project at the Department of Information Studies, University of Sheffield. The rest of the work was funded as part of the EU 5th Framework project, MIND: contract number IST-2000-26061. Further information on the project can be found at http://mind.cis.strath.ac.uk/.

Reference

1. John S. Garofolo, Cedric G. P. Auzanne, Ellen M. Voorhees; "The TREC Spoken Document Retrieval Track: A Success Story"; Proceeding of the 8th Text REtrieval Conference (TREC 8), E. Voorhees, Ed.; Gaithersburg, Maryland, USA; November 16-19, 1999; pp107
2. D. Abberley, S. Renals and G. Cook; "Retrieval of broadcast news documents with the THISL system"; In Proceeding IEEE ICASSP, pp 3781-3784; Seattle, 1998
3. John S. Garofolo, Ellen M. Voorhees, Vinvent M. Stanford, Laren Sparck Jones; "TREC-6 1997 Spoken Document Retrieval Track Overview and Results"; Proceeding of the 6th Text REtrieval Conference (TREC 6); November 19-21, 1997; pp83
4. M.A. Siegler, M.J. Witbrock, S.T. Slattery, K. Seymore, R.E. Jones and A.G. Hauptmann; "Experiments in Spoken Document Retrieval at CMU"; Proceeding of the 6th Text REtrieval Conference (TREC 6); November 19-21, 1997; pp291-302
5. J.S. Garofolo, E.M. Voorhees, C.G.P. Auzanne, M. Stanford, B.A. Lund; "1998 TREC-7 Spoken Document Retrieval Track Overview and Results", in Proceedings of the DARPA Broadcast News Workshop, 1999
6. S.E. Johnson, P. Jourlin, G.L. Moore, K. Spärck Jones, P.C. Woodland; "Spoken Document Retrieval For TREC-7 At Cambridge University"; Proceeding of the 7th Text REtrieval Conference (TREC 7); pp191
7. Singhal, J. Choi, D. Hindle, D.D. Lewis, F. Pereira; "AT&T at TREC-7"; Proceeding of the 7th Text REtrieval Conference (TREC 7); pp239
8. P. Nowell; "Experiments in Spoken Document Retrieval at DERA-SRU"; Proceeding of the 7th Text REtrieval Conference (TREC 7); pp353
9. Allan, J. Callan, M. Sanderson, J. Xu; "INQUERY and TREC-7"; Proceeding of the 7th Text REtrieval Conference (TREC 7); pp201
10. Jean-Manuel Van Thong, David Goddeau, Anna Litvinova, Beth Logan, Pedro Moreno and Michael Swain; "SpeechBot: A Speech Recognition based Audio Indexing System for the Web"; International Conference on Computer-Assisted Information Retrieval, Recherche d'Informations Assistee par Ordinateur (RIAO2000); Paris, April 2000; pp 106-115
11. Pedro Moreno, JM Van Thong, Beth Logan, Blair Fidler, Katrina Maffey,and Matthew Moores; "SpeechBot: A Content-based Search Index for

Multimedia on the Web"; First IEEE Pacific-Rim Conference on Multimedia, (IEEE-PCM 2000), 2000
12. Mark Sanderson and Fabio Crestani; "Mixing and Merging for Spoken Document Retrieval"; Proceedings of the 2nd European Conference on Digital Libraries; Heraklion, Greece, September 1998, pp397-407. Lecture Notes in Computer Science N. 1513, Springer Verlag, Berlin, Germany.
13. Mark Sanderson, Xiao Mang Shou; Speech and Hand Transcribed Retrieval; LNCS 2273, Information Retrieval techniques for Speech Application, Springer, 2002, pp78-85
14. Nick Tuff; MSc dissertation; Department of Information Studies, University of Sheffield; 2002
15. Anastasios Tombros and Mark Sanderson; "Advantages of query biased summaries in information retrieval", Proceedings of ACM SIGIR, 1998, pp2-10
16. Compaq Computer Corporation Cambridge Research Laboratory "The First Internet Site for Content-Based Indexing of Streaming Spoken Audio", Technical white paper, available from http://speechbot.research.compaq.com/.