Speech and Hand Transcribed Retrieval

Mark Sanderson
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, S10 2TN, UK
+44 114 22 22648
m.sanderson@shef.ac.uk

Xiao Mang Shou
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, S10 2TN, UK
+44 114 22 22675
x.m.shou@shef.ac.uk

ABSTRACT

This paper describes the issues and preliminary work involved in the creation of an information retrieval system that will manage the retrieval from collections composed of both speech recognised and ordinary text documents. In previous work, it has been shown that because of recognition errors, ordinary documents are generally retrieved in preference to recognised ones. Means of correcting or eliminating the observed bias is the subject of this paper. Initial ideas and some preliminary results are presented.

General Terms

Measurement, Experimentation.

Keywords

Information Retrieval, Spoken Document Retrieval, Mixed Collections.

1. Introduction

At present, the main stream of information retrieval research remains in homogenous data collections. However, with the progress of multimedia technologies, the trend for searching heterogeneous data collections is growing. The MIND (Multimedia International Digital Libraries) project [MIND01] that the authors are currently working with provides a typical example for this case. The research in MIND addresses issues that arise when people have remote access to a great many heterogeneous and distributed multimedia digital libraries containing text, images, and audio data such as speech recognised documents. In order to cope effectively with such masses of knowledge, a subset of the digital libraries that most likely contains relevant documents must be identified, selected, and searched. The results obtained from those resources need to be fused and merged into a single result. Both the means of resource selection and data fusion must be achieved without bias.

When ranking documents in a collection in relation to a query, almost all of the weighting schemes and retrieval models used assume that documents are created equal. The presence or absence of a query word or the number of times that word occurs within documents is assumed to have the same significance across the collection (once document length has been taken into account). If, however, a collection is composed of documents created in different ways, such as a collection of ordinary and speech recognised documents, or collections of speech recognised documents where the recognition accuracy varies widely, the assumptions are likely to be false.

In general, retrieval of spoken documents is viewed as a success [Garofolo99b], speech recognition (SR) systems are accurate enough to allow retrieval to operate at an acceptable level. Spoken document retrieval (SDR) is being used: internally within corporations ([Abberley98], [Renals99], [THISL01]), and as a Web search engine ([SpeechB00A], [SpeechB00B], [SpeechB01]). Word error rates (WER) on the audio documents being retrieved are still relatively high, however, because relevant documents generally contain each query term with a high term frequency (*tf*), as long as a few of the term occurrences are recognised, relevant documents will be retrieved. If, however, the collection is composed of both speech recognised and ordinary text documents, because the *tf* in the recognised documents is lower, the ordinary documents will be retrieved in preference to the recognised. This problem is also found within the collection of spoken documents, some audio documents are likely to be recognised more accurately than others, those recognised well will contain query terms with higher *tf* than those recognised badly[1].

Within the framework of the MIND project, it is also assumed that remote digital library providers will be either uninterested or unwilling to provide detailed data on the content of their libraries, therefore, it is desirable to determine the information independently. This paper proposes an approach for dealing with this problem (for speech data) through the automatic identification of spoken documents and subsequent estimation of the word error rate within those documents. Once such information has been gained, document retrieval scores will be adjusted to ensure a fairer ranking. The rest of this paper describes past work in this area, followed by a description of the methods to be used in the system. Some preliminary work is then outlined before finally concluding.

2. Past work

Although a great deal of research has been conducted in the retrieval of corrupted data, be it scanned text [Jones01], speech ([Garofolo99b], [Abberley98]), or translated foreign language

[1] For such documents, there is also the problem of no query term occurrences being recognised at all. Clearly an important problem, which has been addressed in past work (e.g. the document expansion of Singhal [Singhal99] and use of multiple hypotheses by Siegler [Siegler97], it is not discussed here.

documents [Franz98], relatively little work has investigated the notion of varying levels of corruption in the collection(s) being retrieved. This is perhaps surprising, as it is quite reasonable to expect such variation. Sanderson and Crestani briefly addressed the problem of retrieving from a collection composed of both ordinary and spoken documents [Sanderson98], reporting that using a conventional *tf·idf* weighting scheme, ordinary documents were retrieved in preference to spoken documents. Since then, research in the area has come from the Cross Language Information Retrieval (CLIR) community where retrieval from collections composed of documents written in different languages requires consideration of translation accuracy. Poorly translated documents are less likely to be retrieved. Effective methods to deal with this situation have been devised by identifying the language of the documents and biasing ranking algorithms based on training data derived from cross language test collections [Franz98]. Reported as being effective, the only problem with this technique is the assumption that corruption from translation error is consistent across documents of the same language, an assumption which is likely to be false with the corruption from recognition used in MIND.

3. Methods

The proposed means of providing a fair ranking when dealing with collections of documents composed of either ordinary and spoken documents or spoken document recognised at varying levels of accuracy requires two components: a spoken document identification system; and a method to estimate word error rates. They are described here.

3.1 Spoken document identification

When considering how to spot if a piece of text was written or automatically recognised, one can view spoken document identification as a form of language identification ([LangID01A], [LangID01B], which is used to identify the kind of language that a given document was written in. Though techniques applied in language identification such as training a word or character based n-gram model for recognised output and ordinary text, etc. could also be used, some surface cues in the document are likely to be able to identify whether or not it is the output of speech recognition. For example, speech recognisers generally do not insert punctuation in a fine granularity level and do not have clear boundary between sentences and paragraphs (here only the raw speech recognised data with no human corrections is considered). Therefore, by examining the ratio of punctuation to non-punctuation tokens, and/or the ratio of capitalised to non-capitalised letters may provide sufficient clues for the identification task. Another possible solution could be to check whether a document contains common spelling mistakes as speech recognisers working from a pre-defined vocabulary do not misspell words. Alternatively, a more sophisticated approach is to determine if the document contains any words that are outside the recogniser's lexicon. However, since different speech recognisers have different lexicons and getting the lexicons of them can be difficult, this method cannot be generally applied.

Table 1 shows some preliminary results by comparing the punctuation and upper case words in 1804 speech recognised and 115 hand transcribed files from TREC-9 collections.

From the table, it shows a magnitude difference in the percentage of punctuation in speech recognised and hand transcribed collections. The punctuation within speech documents contain full stops and apostrophe marks only whereas hand transcriptions contains evenly distributed comma, hyphen, parenthesis and question marks as well as full stops and apostrophes. By further examining the speech recognised documents, most of the full stops are not addressed to finish a sentence but used as abbreviations such as H.I.V., U.N., etc. Though strictly they should not be counted as full stops, the magnitude difference in the punctuation percentage is not affected by including them.

The speech recognised documents used in the experiment are the ".srt" format files where all words are in capital letters. This may not always be true as some other speech recognised files may contain lower case letters only. Therefore, our experiment does not cover all situations but extremely high or low rate of upper case words can be used as an indication for speech document identification.

3.2 Word error rate estimation

Once a document has been identified as one generated by a SR, the next stage will be determining how much error is in the transcript. In speech recognition, the conventional method is to measure the Word Error Rate (WER) by comparing the output of the recogniser to an accurate hand transcription. In the situation we anticipate such a transcript will not be available. In seeking an automatic method for estimating error rates, one should consider how speech recognisers determine the most likely sequence of words that were spoken. In addition to the acoustic models that recognise phonemes, modern SR systems use language models derived from large corpora to determine the most likely word sequence given a set of recognised phonemes. The models are applied over a small sliding window of uttered words (no more than four or five) and have the effect of improving recognition quality. However, even when incorrectly recognising words, the models produce texts that at the level of word pairs or triples make some form of sense. For example the word sequence "On world news tonight this Wednesday..." is recognised by a poor speech recogniser as "on world is unlike his wins the...": nonsensical, but each word pairing "on world", "world is", "is unlike", "unlike his", "his wins" & "wins the" are sensible when viewed on their own.

	Total words	**Total punctuation**	**Percent of punctuation**	**Total upper case words**	**Percent of upper case words**
Speech recognised files	9,117,820	360,359	0.04	9,117,820	1.00
Hand transcriptions	822,760	198,151	0.24	59,401	0.07

Table 1: Comparison of punctuation and upper case words in speech recognised and hand transcribed files

Consequently, searching a recognised text for unusual word pairs or triples to estimate WER is unlikely to be fruitful. It is better to examine attributes of documents beyond local context.

The method we propose is based on the observation made in the retrieval experiments of previous work by one of the authors (i.e. [Sanderson98]): where ordinary documents were found to be retrieved in preference to the speech recognised. The reason for this was the lower *tf* value of query terms in the recognised documents when compared to the hand transcribed. This was caused by the SR system failing to recognise all occurrences of all the document words and replacing the mis-recognised words with others. We postulate that, in general, the poorer the recognition accuracy of an audio sequence, the greater the number of single occurrence words within the transcription produced.

To test this, a small initial experiment was conducted using the TREC SDR 1998 collection: a collection composed of approximately 3,000 short documents transcribed from US news reports. The average number of occurrences of all words in a document was computed and the average of this number was calculated across all documents of the collection. The resulting single value provided a simple measure of how often words are repeated within collection documents. Thanks to efforts by SDR track participants in 1998, transcripts from eight SR runs were made available to fellow participants along with the hand transcription, all shared runs were used in the experiment. As reported by Garofolo et al [Garofolo99a], WERs on the runs varied greatly: the runs plus their error rates are listed in table 2 ranked by WER.

The average term occurrence measure (*to*) for each of the transcripts along with the score for the hand transcript was computed. The results are shown in table 3 with the transcripts ordered by score. As can be seen even in the hand transcription words are rarely used more than once, however the score with one notable exception (i.e. NIST B2) does appear to be ranking systems by their WER.

A closer examination of the NIST B2 transcripts indicated that the reason for the high term occurrence measure was due to words with low *idf* being repeated many times within documents. Consequently the experiment was repeated but with the 1,000 highest frequency words ignored when measuring. The results of this experiment are shown in table 4 below. (Note, the overall occurrence rate measure is lower for all transcripts, this is to be expected given the removal of the more frequent terms.) Using a simple correlation test, the ranking produced by the altered measure was closer to the

Rank	System	Description	WER (%)
1	LTT	Hand transcription of audio data	0.0
2	CU HTK	Cambridge University using HTK toolkit	24.6
3	Dragon	University of Massachusetts & Dragon systems	29.5
4	ATT	AT&T	31.0
5	NIST B1	NIST B1 system	33.8
6	Shef	Sheffield University using Abbot system	35.6
7	NIST B2	NIST B2 system	47.5
8	DERA S2	Defence Evaluation and Research Agency, UK	61.3
9	DERA S1	Defence Evaluation and Research Agency, UK	66.0

Table 2: Description and WER of SDR 1998 SR systems.

Rank	System	*to*	WER Rank
1	LTT	1.28	1
2	CU HTK	1.21	2
3	ATT	1.19	4
4	NIST B1	1.19	5
5	NIST B2	1.18	7
6	Dragon	1.18	3
7	Shef	1.18	6
8	DERA S2	1.13	8
9	DERA S1	1.12	9

Table 3: SR systems ranked by term occurrence measure

Rank	System	*to*	WER Rank
1	LTT	1.18	1
2	CU HTK	1.12	2
3	NIST B1	1.11	5
4	ATT	1.11	4
5	Dragon	1.10	3
6	Shef	1.09	6
7	NIST B2	1.07	7
8	DERA S2	1.06	8
9	DERA S1	1.05	9

Table 4: Systems ranked by modified occurrence measure

WER rate ranking in Garofolo et al's paper.

4. Conclusions and future work

This paper has described in part, the retrieval problem of the MIND project and described initial work tackling the important problem of retrieving from collections of documents with varying levels of error within them. A short experiment showed that the proposed ideas are promising and deserving of further investigation and it must be remembered that the method illustrated here only provides a means of ranking transcripts against each other. What is desired is a means of determining levels of error rate from a single transcript. How to achieve this, is our next challenge.

5. Acknowledgements

This work was funded by the EU 5th Framework project, MIND: IST-2000-26061 (mind.cs.strath.ac.uk).

6. References

[Abberley98] 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

[Franz98] M. Franz, J.S. McCarley, S. Roukos; "Ad hoc and Multilingual Information Retrieval at IBM"; Proceeding of The Seventh Text REtrieval Conference (TREC 7); November 9-11, 1998; pp157-168

[Garofolo99a] 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

[Garofolo99b] John S. Garofolo, Cedric G. P. Auzanne, Ellen M. Voorhees; "The TREC Spoken Document Retrieval Track: A Success Story"; Text Retrieval Conference (TREC) 8, E. Voorhees, Ed.; Gaithersburg, Maryland, USA; November 16-19, 1999

[Jones01] G.J.F.Jones and M.Han; "Retrieving Scanned Documents from a Mixed-Media Document Collection"; Proceedings of the BCS-IRSG European Colloquium on IR Research; Darmstadt, Germany, pp136-149, April 2001

[LangID01A] Automatic Language Identification Bibliography; http://speech.inesc.pt/~dcaseiro/html/bibliografia.html, last accessed: June 2001

[LangID01B] Language Identification; http://translation-guide.com/language_identification.htm, last accessed: June 2001

[MIND01] Resource Selection and Data Fusion for Multimedia International Digital Libraries;

http://mind.cs.strath.ac.uk/, last accessed: June 2001

[Renals99] Steve Renals and Dave Abberley; "The THISL SDR system at TREC-9"; Proceedings of TREC-9; http://www.dcs.shef.ac.uk/~sjr/pubs/2001/trec9.html, last accessed: June 2001

[Sanderson98] 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.

[Siegler97] 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

[Singhal99] Amit Singhal, Fernando C. N. Pereira; "Document Expansion for Speech Retrieval"; SIGIR 1999; pp34-41

[SpeechB00A] 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

[SpeechB00B] 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

[SpeechB01] Compaq Speechbot;

http://www.compaq.com/speechbot/, last accessed: June, 2001

[THISL01] The THISL Spoken Document Retrieval Project;

http://www.dcs.shef.ac.uk/spandh/projects/thisl/overview-oct98/, last accessed: June, 2001