ACM SIGIR 2001 Workshop
"Information Retrieval Techniques for Speech Applications"

Anni R. Coden, Eric Brown
IBM T.J. Watson Research Center
{anni, ewb}@us.ibm.com

Savitha Srinivasan
IBM Almaden Research Center
savitha@almaden.ibm.com

ABSTRACT

We organized a workshop at SIGIR'01 to explore the area of information retrieval techniques for speech applications. Here we summarize the results of that workshop

1. INTRODUCTION

Interest in speech applications dates back several decades. However, it is only in the last few years that automatic speech recognition has left the confines of the basic research lab and become a viable commercial application. Speech recognition technology has now matured to the point where speech can be used to interact with automated phone systems, control computer programs, and even create memos and documents. Moving beyond computer control and dictation, speech recognition has the potential to dramatically change the way we create, capture, and store knowledge. Advances in speech recognition technology combined with ever decreasing storage costs and processors that double in power every eighteen months have set the stage for a whole new era of applications that treat speech in the same way that we currently treat text [4, 13].

This workshop was held on September 13th, 2001 in New Orleans, USA as part of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. The goal of this workshop was to explore the technical issues involved in applying information retrieval and text analysis technologies in the new application domains enabled by automatic speech recognition. These possibilities bring with them a number of issues, questions, and problems. Speech-based user interfaces create different expectations for the end user, which in turn places different demands on the back-end systems that must interact with the user and interpret the user's commands. Speech recognition will never be perfect, so analyses applied to the resulting transcripts must be robust in the face of recognition errors. The ability to capture speech and apply speech recognition on smaller, more powerful, pervasive devices suggests that text analysis and mining technologies can be applied in new domains never before considered.

This workshop explored techniques in information retrieval and text analysis that meet the challenges in the new application domains enabled by automatic speech recognition. We posed seven questions to the contributors to focus on:

1. What new IR related applications, problems, or opportunities are created by effective, real-time speech recognition?
2. To what extent are information retrieval methods that work on perfect text applicable to imperfect speech transcript?
3. What additional data representations from a speech engine may be exploited by applications?
4. Does domain knowledge (context/voice-id) help and can it be automatically deduced?
5. Can some of the techniques explored be beneficial in a standard IR application?
6. What constraints are imposed by real time speech applications?
7. Case studies of specific speech applications - either successful or not.

The contributions covered a wide spectrum of topics, which are summarized in the sections that follow.

2. TRADITIONAL INFORMATION RETRIEVAL TECHNIQUES

The paper by James Allan [1] (the keynote speaker) reviews the effects of errors in speech transcripts on traditional information retrieval tasks and shows that these effects are minimal. Allan points out that the tasks studied are based on the traditional information retrieval frame-

work and, in particular, the queries are longer than those typically seen in web or emerging speech applications. Furthermore, the evaluation is based on returning documents instead of paragraphs, sentences or answers. Allan speculates on how a different framework might change the results and supposes that the effects of speech recognition errors will be more noticeable for short queries. He claims that one of the big challenges ahead is how to design user interfaces for the new applications being developed based on speech instead of clean text.

Allan remarks also that a synergy between language models for ASR systems and IR systems may prove helpful in new applications. One of the challenges in the future is how to carry out IR tasks on "mixed" collections of documents.

3. SPOKEN DOCUMENT PRE-PROCESSING

The first area to consider is the collection itself, which traditionally contained only textual documents. However, multimedia collections are becoming more prominent and one of the media is speech. Collections may contain speech recordings or transcripts of speech recordings and this new mixed environment poses new challenges.

Speech recordings are either translated into text (i.e., the transcript), or the application tasks are performed directly on the recordings. Examining the transcripts, they are in general not perfect and tools have been developed which improve their quality. An example of a novel tool is the automatic capitalization of mono-case text. This is an important task as closed caption text for instance is mono-case. Furthermore, automatically generated transcripts could have the wrong capitalization and it could prove important that capitalization be uniform. Brown and Coden [2] present a discussion of these issues and several algorithms in the paper "Capitalization Recovery for Text". Their approach is based on a training corpus from which a set of dictionaries of named entities and phrases is build. To recover capitalization from case deficient text, several processing steps including heuristics and dictionary lookup are performed. Initial accuracy results are quite encouraging.

4. ADAPTING IR TECHNIQUES TO SPOKEN DOCUMENTS

Many known information retrieval tasks traditionally operate on well edited and grammatically correct text. It is an area of research to explore how these tasks operate on speech transcripts. An alternate approach is to execute the same IR tasks directly on speech recordings and compare the performance both in terms of effectiveness of the execution and compute power. The paper "Clustering of Imperfect Transcripts using a Novel Similarity Measure" by Ibrahimov, Sethi and Dimitrova [9] discusses a traditional IR task (*clustering*) on a novel domain of data (speech transcripts). They propose a novel similarity measure designed for corpora transcribed by an Automatic Speech Recognition engine. Desilets, de Bruijn and Martin [6] explore another traditional IR task – *keyphrase extraction* – on this novel domain in the paper "Extracting keyphrases from Spoken Audio Documents". They show that keyphrase extraction algorithms seem to be robust for corpora with non-zero word error rates. Several experiments are discussed. It was stipulated that ASR systems might be able to recognize long words better than short ones.

A traditional IR task is to *segment text* and the paper "Segmenting Conversations by Topic, Initiative and Style" by Ries [11] explores speech features for the segmentation process. In particular, he presents a probabilistic segmentation algorithm based on these features and shows good results. The paper "Extracting Caller Information from Voicemail" by Huang et al. [8] describes a traditional IR task (*extracting information*) in a spoken word environment. The information needed is very specific to this environment, for instance, extracting all names or all numbers would not satisfy the goal of the application. The task is to detect the correct subset. Furthermore, a named entity in a phone conversation may be the whole phrase "Mark from purchasing," as opposed to just "Mark," and a phone number may be comprised of digits and words like "extension". They developed a maximum entropy model that performed well but degraded significantly with high word error rates.

5. TECHNIQUES FOR MULTI-MEDIA COLLECTIONS

To date, most tasks are performed either on a purely textual collection or solely on speech recordings or their transcripts. Tasks applied to mixed collections seem to favor textual data, which may not be always appropriate. The paper "Speech and Hand Transcribed Retrieval" by Sanderson and Shou [12] explores this area. In particular, it explores whether one could apply traditional IR techniques with varying parameters to address the different types of data in the collection. In had been previously shown, that hand transcribed documents rank higher in search results than documents which have been automatically transcribed. To rectify this problem, workshop participants discussed a novel ranking procedure for mixed collections. In particular, a collection would be split into multiple uniform sub-collections, all documents in a sub-collection being either hand or automatically transcribed. Search would be carried out on each collection separately and the results merged in an unbiased way. To carry out such an approach, methods for determining the modality of a document have to found. Furthermore, it was discussed that such methods may have to operate in real-time. It was suggested that the confidence levels returned by speech recognizers should be taken into account.

6. NEW APPLICATIONS

So far, the focus was on traditional information retrieval tasks where the data on which these tasks are performed is speech. However, with a new domain, new tasks can be exploited. First, it has to be investigated whether users have different expectations when performing tasks on speech. The paper by Kim and Oard [10] entitled "The Use of Speech Retrieval Systems: A Study Design", proposes a user study whose goal is to determine the *user expectation* in this emerging field. How is information accessed and how is it used? A potential outcome of the study would be to determine the features of a document that make it relevant to a users search. Such features could then be the meta-data associated with a document.

The paper "Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition" by Fujii, Itou and Ishikawa [7] addresses the issues associated with using speech as input to an information retrieval task and show several results in this area (*integration, adaptation).* They propose to use the target text collection to improve the statistical language model for the automatic speech recognition engine and preliminary experiments warrant more research in this area.

Another aspect of the user expectations that has a huge impact on all parts of an application is a "*real-time*" requirement. In every day life, people expect an immediate response during discourse. The work by Coden and Brown [3, 5, 4] focuses on the real-time aspect and the papers on Data BroadCasting and Meeting Mining address this issue within a comprehensive framework and architecture. The Data BroadCasting works focuses on automatically inserting relevant additional data into a live television broadcast data stream. The Meeting Mining application aims at enriching (in real-time) a meeting by retrieving appropriate data ranging from the answer to a question to presenting a memo a participant mentioned.

7. CONCLUSIONS

The desire for more information is increasing with the amount of data stored and accessible to the general population. The data currently stored is not only text documents, but includes other media like speech, video and images. The focus of this workshop was on speech applications. Several papers showed that traditional IR tasks perform reasonably well on ASR transcribed documents. One of the challenges going forward is how to deal with these tasks on non-uniform collections, some of the documents being ASR transcribed and others being written and more grammatically correct. In particular:

How does one determine the modality of a document?

How are ranking/weighting/scoring schemata adjusted for various modalities?

How are standard IR tasks carried out on non-uniform collections?

What new applications are possible based on a robust ASR system? Data Broadcasting and MeetingMiner were presented as examples of information retrieved automatically by "understanding" the spoken word. These applications

are in their infancies but point to the future. They also exemplify a different user expectation – the tasks have to be carried out in real time.

The user expectations for new applications were discussed and it became clear that a methodology and framework has to be developed to evaluate these. Work is underway in this area.

There seems to be a lot of interest and work in the community on developing new information retrieval applications based on multi-media data. Some of the follow-up events from this workshop are: A tutorial at SIGIR2002 organized by D. Ponceleon and S. Srinivasan and a AAAI Spring 2003 symposium organized by Gareth Jones, Ruud Bolle, Anni Coden, Alex Hauptmann, Corinna Ng and Shin'ichi Satoh.

The papers presented at the workshop where published by Springer Verlag in the Lecture Notes in Computer Science series (LNCS 2273) and this article is based on the preface of this book.

8. REFERENCES

[1] J. Allan, Perspectives on Information Retrieval and Speech, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[2] E. W. Brown and A. R. Coden, Capitalization Recovery for Text, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[3] E. W. Brown and A. R. Coden, WASABI: Framework for Real-Time Speech Analysis Applications, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[4] E. W. Brown, S. Srinivasan, A. Coden, D. Ponceleon, J. W. Cooper and A. Amir, Toward Speech as a Knowledge Resource, IBM Systems Journal, 40 (2001), pp. 985-1001.

[5] A. Coden and E. Brown, Speech Transcript Analysis for Automatic Search, Proc. of HICSS'34, 2001.

[6] A. Desilets, B. d. Bruijn and J. Martin, Extracting Keyphrases from Spoken Audio Documents, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[7] A. Fujii, K. Itou and T. Ishikawa, Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[8] J. Huang, G. Zweig and M. Padmanabhan, Extracting Caller Information from Voicemail, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[9] O. Ibrahimov, I. Sethi and N. Dimitrova, Clustering of Imperfect Transcripts Using a Novel Similarity Measure, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[10] J. Kim and D. W. Oard, The Use of Speech Retrieval Systems: A Study Design, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[11] K. Ries, Segmenting Conversations by Topic, Initiative, and Style, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[12] M. Sanderson and X. M. Shou, Speech and Hand Transcribed Retrieval, Proc. of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, Springer LNCS 2273, 2002.

[13] S. Srinivasan and E. Brown, Is Speech Recognition Becoming Mainstream, IEEE Computer, 35 (2002), pp. 38-41.