Linking Transcribed Conversational Speech

Joseph Malionek
Eleanor Roosevelt High School
Greenbelt, MD USA
jmalionek@gmail.com

Douglas W. Oard
University of Maryland
College Park, MD USA
oard@umd.edu

Abhijeet Sangwan
John H.L. Hansen
The University of Texas at Dallas, USA
abhijeet.sangwan@utdallas.edu
john.hansen@utdallas.edu

ABSTRACT
As large collections of historically significant recorded speech become increasingly available, scholars are faced with the challenge of making sense of what they hear. This paper proposes automatically linking conversational speech to related resources as one way of supporting that sense-making task. Experiment results with transcribed conversations suggest that this kind of linking has promise for helping to contextualize recordings of detail-oriented conversations, and that simple sliding-window bag-of-words techniques can identify some useful links.

Categories and Subject Descriptors
H.3.m [**Information Systems**]: Information Storage and Retrieval – *miscellaneous.*

General Terms
Experimentation.

Keywords
Content linking, conversational speech.

1. INTRODUCTION
Dramatic reductions in the cost of audio recording in the 1960's yielded a transformation in what was recorded. Before that time, most recorded speech was "formal," in the sense that it was produced to convey information to some audience. Typical examples of recorded speech from that time include radio broadcasting and political speeches. Starting in the 1960's, however, it became increasingly common to record conversations. Prominent examples from that time that are of interest to scholars today include President Johnson's recorded telephone calls, President Nixon's recorded meetings, and NASA's recorded radio conversations with astronauts on the Moon.

Listening in on other people's conversations poses several challenges, however. One problem is how to know which parts of a large collection are worth listening to; that's the well-researched speech retrieval problem. Another problem is understanding what you are hearing; "insider language" that a non-participant might not easily understand without access to the broader content of the recorded interaction is typically laced throughout conversations that are incidentally recorded. A third problem is that task-focused conversations are often incomplete, since once the task at hand has been dealt with the participants in the conversation don't have any reason to fill out the rest of the story.

In this paper, we begin to explore automated content linking as a potential way of supporting contextualization when listening to a recorded conversation. In our experiments, we start from recorded radio conversations with the Apollo astronauts, and we seek to automatically create links to oral history interviews that were recorded many years later with Apollo program participants. One advantage of this experimental setting is that transcripts are already available for both the Apollo radio conversations and the Apollo oral history interviews. Another reason for our choice of this setting is that we have already built a system for synchronized replay of the radio audio (along with photographs, video, and other materials from the missions), which in future work we expect will facilitate usability studies that will allow us to characterize the actual utility of automatically building such links.

The remainder of this paper is organized as follows. We begin by reviewing related work on automated content linking, searching conversational speech, and linking conversational speech. Section 3 then describes our experiments, including the collections that we linked, our approach to building links automatically, our evaluation design, and the evaluation results. Section 4 concludes the paper with a few brief remarks on our planned next steps.

2. RELATED WORK
Research on automated linking is not new. As one example, the Story Link Detection task in the Topic Detection and Tracking evaluations involved linking entire news stories, including transcribed news broadcasts [1]. News broadcasts consist principally (but not exclusively) of planned (and, thus, formal) speech, however, whereas our focus is on conversational speech.

The Initiative for Evaluation of XML Retrieval (INEX) Link-the-Wiki task extended the focus of content linking to include pinpointing the span in a document from which a link should be built [2]. The key idea in the Link-the-Wiki task was to use the links already present in Wikipedia pages as "found data" for training and evaluating automated linking techniques. Some approaches developed for this task proved to be extensible to other sources (e.g., news articles [3]), although the most effective techniques relied on structural characteristics of Wikipedia as a target. To the best of our knowledge, our focus on pinpointing on both the source and target sides of the link is novel.

In 2012, the Forum for Information Retrieval Evaluation (FIRE) Cross-Language Indian News Search (CL!NSS) task extended the focus on linking in another way, to linking documents that were represented in different feature spaces (in this case, in different languages) [4]. For our experiments we use a consistent feature space for both source and target (we use transcribed speech), but we see our work as preparatory to research in which we will seek to link from untranscribed speech to text sources.

SIGIR '13, July 28 - August 01 2013, Dublin, Ireland

ACM 978-1-4503-2034-4/13/07…$15.00.

Although we are not aware of prior work with linking between different sources of conversational speech, we can build on prior research on ranked retrieval from conversational speech. One line of work involves searching recorded telephone conversations (e.g., [5]). Test collections built from telephone speech have some limitations, however, since privacy considerations limit the distribution of naturally occurring telephone conversations, while the redistributable telephone conversation collections created for speech research include either a relatively small number of easily distinguished topics (e.g., the Switchboard and Fisher corpora) or they include mostly "chit chat" for which construction of realistic topics can be problematic (e.g., Call Home).

An alternative source of conversational speech is oral history interviews, which are somewhat less spontaneous than telephone speech but which do include substantial informal interaction. The Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) evaluation produced two information retrieval test collections, one in English and one in Czech [6]. The Czech collection is notable for our purposes because it requires pinpointing in unsegmented speech (the English was pre-segmented). One limitation of the CLEF CL-SR test collections is that only automatically generated transcripts are available; it can also be useful to have manually prepared transcripts a basis for comparison. In our experiments we work with manually transcribed oral history interviews as the target of our linking task.

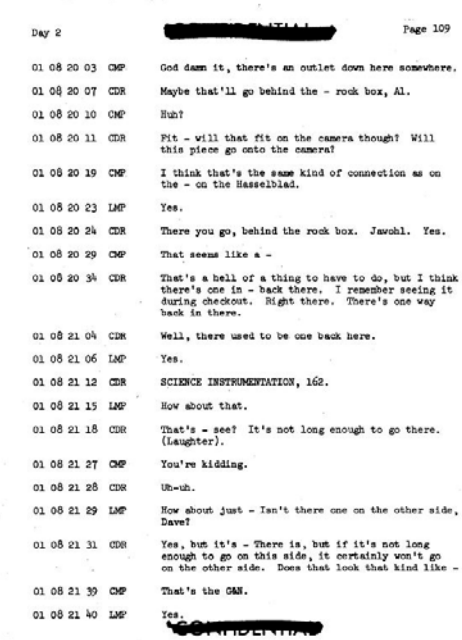
We are aware of one study that sheds some light on how linking conversational speech might be useful, at least in an academic context. In that work, parts of two oral history interviews were manually linked to related Web-accessible resources and then those resources were categorized by type and by purpose [7]. A total of 91 links were made to Wikipedia (24%), other Web Sites (14%), historical newspapers (14%), primary source materials such as personal papers collections or other oral histories (13%), books (11%), and videos (10%) (magazines, maps, interviews, and scholarly publications were each targets for fewer than 10% of the manually created links). The authors' self-report of their motivation for creating links indicated two general reasons: elaboration (79%) or contextualization (21%). In the experiments presented in this paper, we focus on contextualization because we link to, rather than from, oral history interviews.

3. AUTOMATED LINKING EXPERIMENT

Here we describe how we obtained and processed the radio transcripts that we linked from and the oral history transcripts that we linked to. We then describe how we automate the linking process, how we constructed a small test collection for use in our experiments, the design of those experiments, and our results.

3.1 Transcribed Speech Collections

We obtained the transcripts of the radio communication between the Apollo spacecraft and the Mission Control Center in Houston Texas for the Apollo 14 and 15 missions.[1] The transcript for each mission is a single searchable PDF file that had been created by scanning the typewritten transcript that was prepared originally for engineering analysis after completion of each mission. Figure 1 shows an excerpt from the Apollo 15 transcript. We performed page layout analysis and OCR using an integrated analysis system for structured documents [8], and then used positional heuristics to correct some of the misrecognized digits in the Ground Elapsed Time (GET, the number of days, hours, minutes, and seconds since launch, which Figure 1 shows on the left) for each utterance by automatically correcting letters (which cannot appear in a time field) to the most likely digit (e.g., upper case letter O to digit 0, lower case letter l to digit 1, upper case letter S to digit 8) and then enforcing the same strict temporal ordering that had been present in the original on the character-corrected times (this serves to detect some mis-corrections). We then indexed each utterance by GET for use in our mission reconstruction system.


Day 2 [redacted] Page 109

01 08 20 03	CMP	God damn it, there's an outlet down here somewhere.
01 08 20 07	CDR	Maybe that'll go behind the - rock box, Al.
01 08 20 10	CMP	Huh?
01 08 20 11	CDR	Fit - will that fit on the camera though? Will this piece go onto the camera?
01 08 20 19	CMP	I think that's the same kind of connection as on the - on the Hasselblad.
01 08 20 23	LMP	Yes.
01 08 20 24	CDR	There you go, behind the rock box. Jawohl. Yes.
01 08 20 29	CMP	That seems like a -
01 08 20 34	CDR	That's a hell of a thing to have to do, but I think there's one in - back there. I remember seeing it during checkout. Right there. There's one way back in there.
01 08 21 04	CDR	Well, there used to be one back here.
01 08 21 06	LMP	Yes.
01 08 21 12	CDR	SCIENCE INSTRUMENTATION, 162.
01 08 21 15	LMP	How about that.
01 08 21 18	CDR	That's - see? It's not long enough to go there. (Laughter).
01 08 21 27	CMP	You're kidding.
01 08 21 28	CDR	Uh-uh.
01 08 21 29	LMP	How about just - Isn't there one on the other side, Dave?
01 08 21 31	CDR	Yes, but it's - There is, but if it's not long enough to go on this side, it certainly won't go on the other side. Does that look that kind like - -
01 08 21 39	CMP	That's the G&N.
01 08 21 40	LMP	Yes.

[redacted]

Figure 1. A page from the Apollo 15 transcript.

We obtained machine-generated searchable PDF transcripts for 270 oral history interviews that had been conducted with Apollo Program participants by the Johnson Space Center Oral History Project.[2] Figure 2 shows an excerpt from one interview. We extracted the text from each PDF file and then automatically segmented each transcript on the capitalized interviewer name (a reliable transcription convention in this collection to indicate a speaker turn start) to produce question-answer pairs. We then indexed the question, the answer, and the interviewee name as separate fields for a single short "QA triple" using Lucene. We indexed a total of 8,838 QA triples from 170 interviews, an average of about 52 triples per interview.

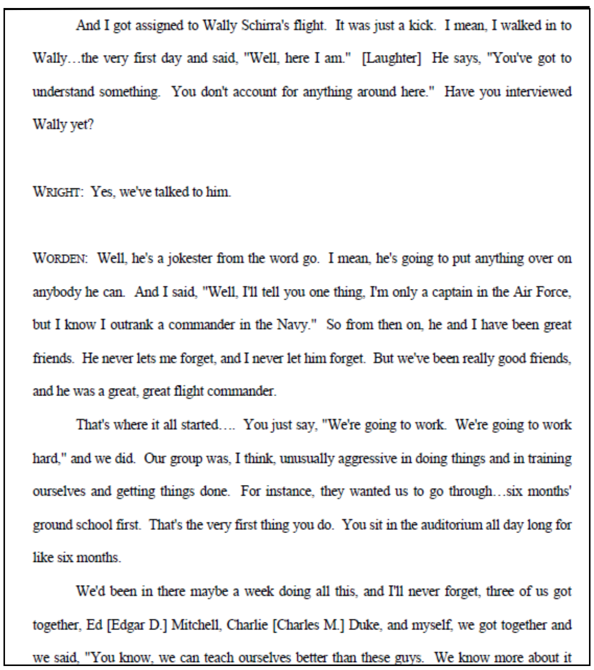
3.2 Automated Linking

Our goal is to automatically link from a time in the mission transcript to a QA triple. QA triples contain an average of 93 words (min 32, max 3,421), which seemed to us to be a reasonable scope for these initial experiments with automatic support for contextualization. We display the first part of three QA pairs (i.e., interviewee name, question, and part of the answer) at each time, and we provide a drill down capability to allow the user to see the full content of any QA pair they wish to select for detailed reading.

[1] http://www.jsc.nasa.gov/history/mission_trans/mission_transcripts.htm

[2] http://www.jsc.nasa.gov/history/oral_histories/oral_histories.htm

In the mission transcripts the transcribed speech is segmented into (usually brief) transmissions that average 16 words (min 1, max 2,533), where a transmission is contiguous speech (i.e., without a long break) by a single speaker. We refer to a transcribed transmission as an "utterance." To represent the content being spoken at a time, we store the entirety of the most recently started utterance. If that utterance contains fewer than some pre-specified number of words (or other tokens), the preceding and following utterances are added to the representation in their entirety. If the result is still too short, the process repeats until the representation contains at least the specified minimum number of words (or other tokens. In our experiments, that minimum number is set to 5, 10, or 20. All non-alphanumeric characters are then removed from the representation, and the resulting representation is used by Lucene as a bag-of-words query to search the QA triples (based only on the text in the question and answer fields, weighted equally (the interviewee name field is not searched).



And I got assigned to Wally Schirra's flight. It was just a kick. I mean, I walked in to Wally…the very first day and said, "Well, here I am." [Laughter] He says, "You've got to understand something. You don't account for anything around here." Have you interviewed Wally yet?

WRIGHT: Yes, we've talked to him.

WORDEN: Well, he's a jokester from the word go. I mean, he's going to put anything over on anybody he can. And I said, "Well, I'll tell you one thing, I'm only a captain in the Air Force, but I know I outrank a commander in the Navy." So from then on, he and I have been great friends. He never lets me forget, and I never let him forget. But we've been really good friends, and he was a great, great flight commander.

That's where it all started…. You just say, "We're going to work. We're going to work hard," and we did. Our group was, I think, unusually aggressive in doing things and in training ourselves and getting things done. For instance, they wanted us to go through…six months' ground school first. That's the very first thing you do. You sit in the auditorium all day long for like six months.

We'd been in there maybe a week doing all this, and I'll never forget, three of us got together, Ed [Edgar D.] Mitchell, Charlie [Charles M.] Duke, and myself, we got together and we said, "You know, we can teach ourselves better than these guys. We know more about it

Figure 2. A portion of an oral history interview.

We have integrated a display of linked QA pairs into our existing mission reconstruction system for Apollo missions [9]. This system presents a time-synchronized replay of recorded audio, mission transcripts (for both radio transmissions and recordings made aboard the spacecraft), video, photographs, maps, event timelines, and flight plans; the goals is to help the user to see events from multiple perspectives as they unfolded at the time. To this we added an option to display three linked QA pairs. Because we update the display each second in our mission reconstruction system, the effect is to display the three top-ranked QA triples each time a new utterance started.

Our initial implementation seemed promising, but we were surprised to see that often the QA triples that were displayed were from unrelated missions. The Apollo lunar missions all followed the same general sequence, and all employed the same general types of equipment and procedures. We hypothesized that it was this similarity between missions that was causing the problem, and we therefore added an optional heuristic that, rather than selecting the three top-ranked QA triples for each utterance, would select the three highest-ranked QA triple from either an astronaut who flew the mission or an astronaut who served as "capsule communicator" (CAPCOM) in Mission Control for that mission if such a QA triple could be found (and the top-ranked utterances from others to fill in the set of three, if necessary). We refer to this as the "filtered" condition.

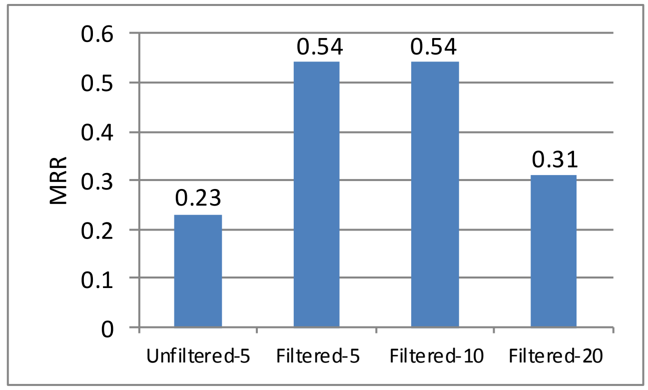


Figure 3. Ranked linking effectiveness results.

3.3 A Test Collection

In order to test our system, we manually created an answer key with known ground truth links. The oral history interviews (which typically lasted a few hours) are far shorter than the missions (which each lasted more than a week), so the choice of events discussed in each oral history interview is necessarily highly selective. We therefore built the answer key by starting with one QA triple from some oral history interview and then manually finding the corresponding time span in the mission transcript. For this initial study, we started only with QA triples from interviews with astronauts who flew one of our two missions (Apollo 14 or 15) or who served as CAPOM for one of those missions.

The first author of this paper built ground truth links in this way for 8 mission events, 4 from Apollo 14 and 4 from Apollo 15. An inter-annotator agreement check by the second author of this paper of two ground truth links found some differences in precise start and end times (e.g., one of us marking the start of a spacewalk when depressurization began, the other one when the hatch was opened after depressurization was completed), but agreement for about 80% of the time span of each event. We therefore designed our evaluation measure to be relatively insensitive to specific start and end times.

3.4 Results

To evaluate our approach to automated linking, we ran our mission reconstruction system over the period indicated in the ground truth and manually noted the highest rank at which the ground truth QA triple appeared in the display during that period. Figure 3 shows the results for four conditions: unfiltered-5 (Lucene query of at least 5 words, no filtering to prefer astronauts who flew or served as CAPCOM on the same mission), filtered-5 (the same query, but with the filter applied), filtered-10 (with a query of at least 10 words), and filter-20 (a query of at least 20 words). We report Mean Reciprocal Rank (MRR) over 8 queries, which awards full credit for the target QA triple in rank 1 at some point during the period, half credit for the target QA triple never higher than rank 2 during the period, and one-third credit for the target QA triple never higher than rank 3 during the period.

Our results show that filtering does appear to improve MRR markedly, but we note that any adverse effect from suppressing QA triples from people who did not fly on or serve as CAPCOMS for the mission would not be seen with our test collection because none of our 8 target QA triples were from an interview with a person outside that set. Nonetheless, our results clearly suggest that it will be important for us to incorporate the missions on which someone worked when building our linking models.

Our results from a sweep across query lengths indicate that shorter queries seem to be preferred, but note that our MRR evaluation measure rewards only the highest position reached by a QA triple, and that the measure is insensitive to any increase in the replacement rate that shorter queries might cause. Ultimately we will want an evaluation measure that rewards a suitable balance between freshness and stability. Moreover, we note that our present approach to query formulation weights all words equally, and that longer queries might prove to be effective if the central terms in those queries (i.e., those uttered closest to the current time) were more highly weighted.

The MRR calculations include reciprocal rank values of zero for two target QA triples that never appeared in the display over the range of times specified in the answer key for any of the four conditions for which we report results. Examining these two consistent failure cases, we found that one was likely missed because the query terms were highly specific words that happened to appear frequently in several other QA triples. In the second case, the same event was mentioned in passing in several different QA triples, and our automated linking approach chose those less appropriate QA triples over the one we had designated where the event was discussed in detail.

In looking at other cases where our system might have done better, we noted a few cases where our system found the right QA triple after the event had ended. Because some complex mission events occur over relatively brief periods, it is common for those events to be discussed retrospectively at some later point in the mission. Because reconstructions of past events can be designed to see into their own future, we might be able to take advantage of this behavior pattern to improve our representation of activities by using the near future as a potential source of expansion text (e.g., by using blind relevance feedback techniques).

4. CONCLUSION AND FUTURE WORK

Our initial results are promising, but much remains to be done, including improving precision and recall, learning when not to make a link, and developing a new (and larger) test collection with ground truth links created by others (because we made the judgments ourselves, the collection we have now is suitable only for development testing). Our early integration with an integrated mission replay system will facilitate studies of how scholars and others will actually employ the capabilities we are creating, which in turn will help us to design evaluation measures that reflect user behavior with higher fidelity.

Looking further to the future, our work with manually prepared transcripts can serve as a baseline for similar work with the actual speech, where transcription errors will naturally be more of a problem. Much of the research on automatic transcription has focused on minimizing the overall word error rate, which is appropriate when the goal is to read the resulting transcript. However, for the linking task, where we have enough context on both sides to perhaps tolerate fairly high word error rates, we may want to tune the transcription system differently. Indeed, we may not want to actually generate transcripts at all---perhaps what we will really want will be lattice matching techniques that can better represent the unresolved uncertainty. Our setting also calls for some novel work on characterizing the acoustic environment (because background sounds change in different phases of a flight) and the communications channel (because as the Earth turns the configuration of the communications system was changing to include different tracking stations) that will be important if we are to build the best possible representations of what was actually said.

On the other hand, the actual recordings are far richer than mere transcripts of what was spoken can capture. In particular, we are interested in looking beyond the spoken terms to see how we can leverage automatic characterization of speaker identity (e.g., manuscript authors have deposited many sets of recorded interviews with NASA, and for some of those sets the metadata describing who was interviewed is incomplete), automatic characterization of acoustic environments (e.g., applying techniques we have reported in [10] to detect acoustic events such as thruster firings that might indicate spacecraft maneuvers), or automatic detection of speaker traits (e.g., stress, or emotion). Quite clearly, this new task, linking conversational speech, and the new types of spoken content with which we are working, can take us in some interesting and important directions.

ACKNOWLEDGMENTS

This material is based upon work supported by NSF under Grant 1218159. Opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] J. Allan (ed.), 2002. *Topic-Detection and Tracking: Event-Based Information Organization*, Springer.

[2] A. Trotman, D. Alexander & S. Geva, 2010. Overview of the INEX 2010 Link the Wiki Track, in *INEX*.

[3] D. Milne & I. Witten, 2008. Learning to Link with Wikipedia, in *CIKM*.

[4] P. Gupta, P. Couugh, P. Rosso and M. Stevenson, 2012. PAN@FIRE: Overview of the Cross-Language Indian News Story Search Task, in *FIRE*.

[5] J. Mamou, D. Carmel & R. Hoory, 2006. Spoken Document Retrieval from Call-Center Conversations, in *SIGIR*.

[6] G. Jones, Y Zhang & D. Oard, 2007. Overview of the CLEF-2007 Cross-Language Speech Retrieval Task, in *CLEF*.

[7] A. Levi & D. Oard, 2012. From Personal Narrative to Collective Memory: Spinning a Web from Oral History, in *XVII International Oral History Conference*.

[8] B. Karagol-Ayan, D. Doermann & B. Dorr, 2003. Acquisition of Bilingual MT Lexicons from OCRed Dictionaries, in *Machine Translation Summit*.

[9] D. Oard & J. Malionek, 2013. The Apollo Archive Explorer, in *JCDL*.

[10] M. Akbacak & J. Hansen, 2007. Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems, *IEEE Trans. on Speech & Audio Processing*, 15(2)465-477.