

Pitt at CLEF05: Data Fusion for Spoken Document Retrieval

Daqing He and Jae-wook Ahn

School of Information Sciences
University of Pittsburgh, Pittsburgh, PA 15260 USA
{dah44, jaa38}@pitt.edu

Abstract. This paper describes an investigation of data fusion techniques for spoken document retrieval. The effectiveness of retrievals solely based on the outputs from automatic speech recognition (ASR) is subject to the recognition errors introduced by the ASR process. This is especially true for retrievals on Malach test collection, whose ASR outputs have average word error rate (WER) of 35%. We explored data fusion techniques for integrating the manually generated metadata information, which is provided for every Malach document, with the ASR outputs. We concentrated our effort on the post-search data fusion techniques, where multiple retrieval results using automatic generated outputs or human metadata were combined. Our initial studies indicated that a simple unweighted combination method (i.e., CombMNZ) that had demonstrated to be useful in written text retrieval environment only generated significant 38% improvement in retrieval effectiveness (measured by Mean Average Precision) for our task by comparing to a simple retrieval baseline where all manual metadata and ASR outputs are put together. This motivated us to explore a more elaborated weighted data fusion model, where the weights are associated with each retrieval results, and can be specified by the users in advance. We also explored multiple iteration of data fusion in our weighted fusion model, and obtained further improvement at 2nd iteration. In total, our best run on data fusion obtained 31% improvement over baseline, and 4% improvement which is a significant difference.

1 Introduction

Spoken documents become more and more popular in people's information seeking activities along with the advance of information technologies, especially the storage and network communication technologies. However, comparing to the studies performed on text base documents, the achievements on retrieving spoken documents are still far less. Recent remarkable advances in the automatic recognition of spontaneous conversational speech makes it even urgent to study effective spoken document retrieval techniques. This is the reason that we participated CLEF Spoken Document Retrieval (SDR) track, and our goal is to leverage technologies developed for text retrieval into retrieving spoken documents.

Retrieving spoken documents from Malach test collection, a test collection developed by University of Maryland for retrieving spontaneous conversational speech [6], poses some interesting challenges. Before Malach collection, there have been several spoken document collections, whose documents are mostly news broadcast stories, help desk telephone calls, and political speeches. The documents in the Malach collection, however, are interviews of Holocaust survivors, who talk about their personal stories. Because of the genre, the topics, and the emotion involved, the average Word Error Rate (WER) in machine transcripts of the documents, an indicator of the quality of the ASR output, is around 35%. This imposes a great difficulty for searching based on these ASR outputs.

Our major interests for this year's experiments, however, lie on the several forms of human generated metadata associated with the spoken documents. For each document, there are a list of person names mentioned in the documents, the human assigned thesaurus keywords and a brief summary in 2-3 sentences written by the human catalogers during their cataloging process.

We view ASR outputs and human generated metadata as two types of information that are complimentary of each other in retrieval process. On the one hand, ASR outputs provide full and detailed information about the content of the documents, which often could not be totally covered by human generated data. On the other hand, human generated metadata provide focused, human-processed, and high quality information that can be relied on for the accuracy of the retrieval. If we can develop a reliable retrieval method that can combine both information into the retrieval process in such way that their complimentary features can be fully explored, the achieved retrieval effectiveness would be greatly superior than that of any one of them. This is the goal of our studies in this year's CLEF-SDR experiments, and two derived research questions are:

1. how the manual metadata and ASR outputs in Malach collection can be integrated for improving the retrieval effectiveness of the final run?
2. what are the parameters that we can utilize to make the data fusion techniques more effective for our task?

In the rest of this report, we will firstly review some existing data fusion methods in Section 2; discuss in detail the experiment settings in Section 3; then talk about the fusion techniques we developed for the CLEF-SDR experiments in Section 4. Finally we will discuss some further studies in Section 5.

2 Data Fusion

In the literature, the techniques for combining multiple queries, document representations or retrieval results is called "data fusion" [5]. It has been an active topic in text retrieval process, and people have developed many techniques for applying fusion techniques in various retrieval applications. Belkin et al. [1] studied pre-search data-fusion approach by progressively combining Boolean query formulations. Lee [5] provided an analysis of multiple post-search data fusion

methods using TREC3 ad hoc retrieval data. Data fusion also has been applied in cross-language information retrieval [3, 2], recommendation systems [7], and many other areas.

In post-search data fusion approaches, to properly merge retrieval results that are commonly ranklist of documents, the score associated with each document has to be normalized within that list. A often used normalization scheme (see Equation (1)) utilizes the maximum and minimum scores of a ranklist (i.e., *MaxScore* and *MinScore*) in the normalization process [5].

$$NormalizedScore = \frac{UnnormalizedScore - MinScore}{MaxScore - MinScore} \quad (1)$$

Fox and Shaw [4] developed several fusion methods for combining multiple evidence, and they named the methods as CombMIN, CombMAX, CombSUM, CombANZ, and CombMNZ (the definitions of them are shown in table 1). Lee [5] studied these methods, and established that CombMNZ is the best among the four in retrieving TREC ad hoc data.

Table 1. Combining functions proposed by Fox and Shaw

CombMIN	minimum of all scores of a document
CombMAX	maximum of all scores of a document
CombSUM	summation of all scores of a document
CombANZ	CombSUM \div number of nonzero scores of a document
CombMNZ	CombSUM \times number of nonzero scores of a document

3 Experiment Settings

3.1 Malach Test Collection

Malach Test Collection was developed by University of Maryland as part of their effort in Malach project [6]. The collection contains about 7800 segments from 300 interviews of Holocaust survivors. All the segments were constructed manually by catalogers. Each segment contains two automatic speech recognition outputs from the ASR system developed by IBM in 2003 and 2004 respectively. The WER of the two outputs are about 40% and 35% respectively. In addition, there are automatically generated thesaurus terms from a system developed at University of Maryland. Each segment also contains a set of human generated data, including person names mentioned in the segment, average 5 thesaurus labels and 3-sentence summaries.

There are total 63 search topics, 38 of which were available for training, and 25 were held as the testing topics. Each topic is designed in TREC style, which has a title, a description and a narrative (see Figure 1). The topics are available

in English, Spanish, Czech, German and French, however, we only used English topics for our studies.

```
<top> <num> 1148
<title> Jewish resistance in Europe
<desc> Provide testimonies or describe actions of Jewish resistance in
Europe before and during the war.
<narr> The relevant material should describe actions of only- or
mostly Jewish resistance in Europe. Both individual and group-based
actions are relevant. Type of actions may include survival (fleeing,
hiding, saving children), testifying (alerting the outside world,
writing, hiding testimonies), fighting (partisans, uprising, political
security) Information about undifferentiated resistance groups is not
relevant.
</top>
```

Fig. 1. An example of the search topic in Malach Collection

3.2 Indri search engine

Our search engine was Indri 1.0, which was a collaboration effort between the University of Massachusetts and Carnegie Mellon University ¹. Its retrieval model is a combination of language model and inference network. We chose it not only because of its state-of-art retrieval effectiveness, but also because it is an open source system, on which we can easily integrated our modifications. Its powerful query syntax is another attraction to us, since we want to specify which index fields should a retrieval be based on for our studies of manual metadata only or automatic data only searches.

3.3 Measures

To study the retrieval results in as wide scenarios as possible, instead of choosing on single measure, we employed a set of evaluation measures, each of which tells us some aspect of the retrieval effectiveness of the search results:

- **mean average precision (MAP)**, the measure aim at giving an emphasis view of precision in a ranklist. Since the ranks of the relevant documents are considered in the measure, this measure gives a reasonable overview of the quality of the ranklist for a given retrieval topic.

¹ <http://newhaven.lti.cs.cmu.edu/indri/>.

- **Precision at 10 (P10)** is a useful measure to examine how much relevant documents are there in the first result screen, which is often the only results viewed by a user.
- **R-Precision (R-PREC)** emphasizes on precision but also avoids the artificial cut-off effect imposed by pre-defined cut-off point, like in P10. The “R” varies according to the number of relevant documents of a given topic.
- **Average Recall at top 1000 returned documents.** It tells the quality of the ranklist from the point of recall.

3.4 Baselines

We established three baselines for evaluating our methods (see Table 2). The first two represent the scenario that no data fusion is performed. We selected a run on ASRTEXT 2004 as the baseline for searching on ASR output, since ASRTEXT 2004 is the better one among the two ASR outputs. This baseline is referred as “*asr04*”, and is treated as the lower baseline. Ideally, we should use the search on manual transcripts as the upper baseline. Since Malach collection does not provide manual transcripts, we used all manually generated data in the segments as a proximate upper bound baseline (referred as “*manual-only*” baseline). We did not apply blind relevance feedback (BRF) over “*manual-only*” run since BRF over “*manual-only*” baseline using Indri’s BRF function generated inferior results. The third baseline is a search on all manual and ASR outputs being put together as if they are different parts of the same document. This represents the simplest data fusion method, and is referred “*simple-fusion*” baseline.

Table 2. Retrieval effectiveness of the three baselines and the CombMNZ run

runs	MAP	R-PREC	P10	Avg-Recall
manual-only	0.2312	0.2836	0.4603	0.5813
asr04	0.0693	0.1139	0.2111	0.3595
simple-fusion	0.1842	0.1985	0.3635	0.5847
autowa1	0.0464	0.0879	0.1683	0.3319
CombMNZ	0.1127	0.1173	0.3079	0.6182

4 Experiments and Results Analysis

4.1 Data Fusion with CombMNZ

The first data fusion method studied in our CLEF-SDR experiments was our implementation of CombMNZ method since Lee demonstrated its superior over the other three methods [5]. This run merged results from three retrieval runs, the “*manual-only*” baseline, the “*asr04*” baseline, and a retrieval run on the

automatic assigned thesaurus keywords called “*AUTOKEYWORD2004A1*” (we call this run “*autokw1*”). Table 2 shows the results of “*CombMNZ*” run and that of the three runs that it was based on. Comparing to the lower “*asr04*” baseline, this combined run has achieved significant improvement by the measures of MAP, P10, and especially Avg-Recall ($P \ll 0.05$ in paired T-tests). However, it generates significant decrease at MAP, R-PREC, and P10 when comparing to the two higher baselines, “*manual-only*” baseline and “*simple-fusion*” baseline ($P \ll 0.05$ in paired T-test). The only improvement it achieved over the two higher baselines is measured by Avg-Recall. This means that “*CombMNZ*” run does return more relevant documents comparing to the two high baselines, but it ranks them badly.

A close exam of the retrieval runs in Table 2 shows that the retrieval effectiveness of “*manual-only*” run is greatly higher than that of the two automatic runs. For example, the MAP of “*manual-only*” increases about 200% over “*asr04*”, the better one of the two automatic runs. Therefore, it makes no sense to assume that their contribution to the final fused ranklist is the same, which is the assumption in CombMNZ model. We need a data fusion model that considers the retrieval difference.

4.2 Weighted CombMNZ

The failure of CombMNZ on our data fusion task motivated us to explore a weighted scheme for data fusion based on CombMNZ. A natural place to insert a weight in CombMNZ is to assign a weight of belief for each retrieval run as it is possible to obtain such evidence or belief. In our weighted CombMNZ model (called WCombMNZ model), such belief is used in calculation of the final combined score for a document (see Equation 2).

$$WCombMNZ_i = \sum_{j=1}^n (w_j \times NormalizedScore_{i,j}) \times n \quad (2)$$

where w_j is a predefined weight associated with a search result to be combined, n is the number of nonzero scores of document i , and the $NormalizedScore_{i,j}$ is calculated using Equation 1.

Table 3. Retrieval effectiveness of individual runs on the 38 training topics

runs	MAP	R-PREC	P10	Avg-Recall
manual-only	0.1494	0.1823	0.3237	0.4221
asr04	0.0525	0.0754	0.1447	0.2788
autowal	0.0239	0.0460	0.0816	0.2832

Various methods can be used to obtain the weight w_j for a given ranklist j . In this year’s experiment, we firstly used the retrieval effectiveness of those

pre-combined runs on the 38 training topics as the weights (the details of the pre-fused runs on the training topics are in Table 3). Therefore, we have four different “*WCombMNZ*” runs (see Table 4), and their retrieval effectiveness evaluated by the four measures are in Table 5.

Table 4. Our weighted combination runs

WCombMNZ-1	use the MAP values as the weights
WCombMNZ-2	use the R-PREC values as the weights
WCombMNZ-3	use the P10 values as the weights
WCombMNZ-4	use the Avg-Recall values as the weights

Table 5. Retrieval effectiveness of The first 4 WCmbMNZ runs on total 63 topics

runs	MAP	R-PREC	P10	Avg-Recall
manual-only	0.2312	0.2836	0.4603	0.5813
CombMNZ	0.1127	0.1173	0.3079	0.6182
WCombMNZ-1	0.2137	0.2589	0.4460	0.6008
WCombMNZ-2	0.1987	0.2431	0.4206	0.6254
WCombMNZ-3	0.1967	0.2416	0.4190	0.6253
WCombMNZ-4	0.1783	0.2188	0.3778	0.6215

All four WCombMNZ runs are significant higher than the non-weighted “*CombMNZ*” run (paired T-test with $P < 0.05$) when looking at MAP, R-PREC, and P10 as the measures. However, they are still significant lower than the “*manual-only*” run using the same measures.

Since the weights of “*WCombMNZ-1*” generated the best MAP, R-PREC and P10 results, we used those weights to help us explore further the effect of different combinations of the weights. As the difference of the retrieval effectiveness between “*manual-only*” run and the two automatic runs is significant higher than that between the two automatic runs, we first explored the change of ratio between the weight of “*manual-only*” run and that of the “*asr-04*” and “*autowa1*” runs. The ratio we tested were 2:1 (that is the weight for “*manual-only*” run is 2, and the weights for the two automatic runs were both assigned to be 1 in WCombMNZ model), 5:1, 10:1, 15:1, and up to 1000:1 (the results are presented in Table 6). We then change the weight ratio between the “*asr04*” and that of “*autowa1*” to 2:1, which is closer to the weight ratio in “*WCombMNZ-1*”, and varied the weight ratio of the three runs from 4:2:1, 6:2:1, and up to 50:2:1. As shown in Table 6, the ratio between the weight of the “*manual-only*” and that of two automatic runs is the dominate factor in affecting the retrieval

performance, and once the ratio between the manual run and the automatic runs is larger than 10, there is not much difference in the retrieval effectiveness evaluated by all measures. However, still none of the fused runs achieves better MAP, R-PREC, and P10 than “*manual-only*”, although they are much closer to the performance of the “*manual-only*” than the two automatic runs, and at the same time, many of them have achieved significant better Avg-Recall than the “*manual-only*” run.

Table 6. Exploring the weight ratios in WCmbMNZ model

runs with ratio	MAP	R-PREC	P10	Avg-Recall	runs with ratio	MAP	R-PREC	P10	Avg-Recall
2-1	0.1884	0.2315	0.4032	0.6236	4-2-1	0.1937	0.2354	0.3735	0.6246
5-1	0.2088	0.2590	0.4254	0.6259	6-2-1	0.2047	0.2523	0.4190	0.6253
10-1	0.2133	0.2598	0.4302	0.6240	10-2-1	0.2110	0.2591	0.4238	0.6236
15-1	0.2132	0.2590	0.4381	0.6211	20-2-1	0.2138	0.2599	0.4333	0.6208
20-1	0.2140	0.2581	0.4413	0.6202	30-2-1	0.2144	0.2591	0.4365	0.6207
25-1	0.2131	0.2581	0.4444	0.6129	50-2-1	0.2141	0.2574	0.4444	0.6172
50-1	0.2141	0.2577	0.4429	0.6133	100-2-1	0.2140	0.2575	0.4444	0.6089
100-1	0.2140	0.2583	0.4460	0.6087					
1000-1	0.2132	0.2593	0.4460	0.5995					

4.3 Multiple Iteration of Data Fusion

One exploration within the data fusion framework is “does multiple iterations of data fusion make sense?” To answer this, we conducted several experiments in WCombMNZ model. Total five retrieval runs were used in the second iteration of data fusion. We kept the “*manual-only*” run since it is the best run so far, and we used the four runs listed in Table 5 “*WCombMNZ-1*” to “*WCombMNZ-4*”. We used the similar scheme to vary the weight ratios among the runs, and we also set all weights to 1 to make WCombMNZ model fall back to CombMNZ so that we can study CombMNZ too. The ratio “2-1” in Table 7 means that the weight for “*manual-only*” is 2, and that for the other four runs is 1.

As shown in Table 7, the “*manual-only*”, in our current retrieval setting, still deserves more weights than the other runs, and the best retrieval results are achieved with the ratio around 10:1. Statistical tests (paired T-test) between the results of the 2nd round fusion runs and that of the “*manual-only*” run demonstrate that all the 2-iteration fusion data generated significant improvement on average recall, but only the runs with ratio above 10:1 generated significant improvement on P10, and only runs with ratio 10:1 and 15:1 generated significant improvement on MAP. No significant improvement can be achieved on R-PREC.

We then generated various 3rd round fusion runs using the similar scheme, which include the “*manual-only*” run, and the four 2nd round runs with the ratio

Table 7. Exploring multiple iterations in data fusion

runs	MAP	R-PREC	P10	Avg-Recall
manual-only	0.2312	0.2836	0.4603	0.5813
2nd-ratio 1-1	0.2119	0.2670	0.4413	0.6241
2nd-ratio 2-1	0.2295	0.2720	0.4540	0.6228
2nd-ratio 5-1	0.2397	0.2806	0.4778	0.6200
2nd-ratio 10-1	0.2409	0.2860	0.4937	0.6188
2nd-ratio 15-1	0.2400	0.2853	0.4968	0.6157
2nd-ratio 20-1	0.2388	0.2856	0.4810	0.6142
3rd-ratio 1-1	0.2403	0.2876	0.5016	0.6143
3rd-ratio 2-1	0.2393	0.2865	0.4857	0.6142
3rd-ratio 1-2	0.2409	0.2866	0.4984	0.6157
3rd-ratio 1-5	0.2407	0.2867	0.4937	0.6159
3rd-ratio 1-10	0.2408	0.2869	0.4921	0.6168
3rd-ratio 1-15	0.2408	0.2864	0.4921	0.6168
3rd-ratio 1-30	0.2404	0.2869	0.4921	0.6171

5:1, 10:1, 15:1 and 20:1. The results are shown in Table 7. None of the 3rd round runs could generate statistical significant improvement over the 2nd round runs. It seems that fusions with iteration more than 2 does not justify the extra costs involved comparing to the 2nd round fusion runs.

Our multiple iteration experiments tell us that it is usually difficult to obtain a better fusion results over the best pre-fusion run when the retrieval effectiveness of the pre-fusion runs are greatly different to each other. The fusion experiment on “*manual-only*” and the other automatic runs is an example of such fusion. However, significant improvement over the best pre-fusion run could be achieved via multiple iterations of fusion. For example, we achieved significant improvement over the best pre-fusion run in two iterations. Of course, we need further experiments to test the general effectiveness of the multiple iteration fusion.

5 Conclusion

In this paper, we have described an investigation of data fusion techniques for spoken document retrieval. Because of the various characteristics of the documents in the Malach test collection, retrieval solely based on the outputs from automatic speech recognition (ASR) is well below retrievals on manual generated data. To overcome the problem, we have explored data fusion techniques for integrating the manually generated metadata information with the ASR outputs. We concentrated on the post-search fusion approach, and explored weighted CombMNZ model with different weight ratios and multiple iterations. Our initial results indicate that a simple unweighted combination method that has demonstrated to be useful in written retrieval environment [5] only generated significant 38% relative decrease in retrieval effectiveness (Mean Average Precision) for our

task by comparing to a simple retrieval baseline where all manual metadata and ASR outputs are put together. Only with the more elaborated weighted combination scheme did we obtained 31% significant relative improvements over the simple fusion baseline, and 4% relative improvement over the manual-only baseline, which is a significant difference.

Our future work include further experiments on the general effectiveness of the multiple iteration fusion, and another future work is to explore the usage of WCombMNZ in other retrieval tasks, where multiple retrieval results can be obtained from one retrieval engine, or even different engines. The third further study we want to work on is to answer the question what is the minimum human generated data to ASR output if the goal is to combine the human generated data with the ASR output to achieve a comparable retrieval effectiveness to a retrieval on manual transcripts.

Acknowledgment

The authors would like to thank Doug Oard, Gareth Jones and Ryen White for their tireless efforts to coordinate CLEF-SDR.

References

1. N.J. Belkin, C. Cool, W.B. Croft, and J.P. Callan. The effect of multiple query representations on information retrieval system performance. In *Proceeding of SIGIR'93*, pages 339–346, 1993.
2. Aitao Chen. Cross-language retrieval experiments at CLEF 2002. In *Proceedings of CLEF 2002*, pages 28–48, 2002.
3. Kareem Darwish and Douglas W. Oard. CLIR Experiments at Maryland for TREC 2002: Evidence Combination for Arabic-English Retrieval. In *Proceedings of TREC 2002*, pages 703–710, 2002.
4. E.A. Fox and J.A. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.
5. Joon Ho Lee. Analyses of multiple evidence combination. In *Proceeding of SIGIR'97*, pages 267–276, 1997.
6. Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, and Samuel Gustman. Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of SIGIR'94*, 2004.
7. Luis M. Rocha. Combination of evidence in recommendation systems characterized by distance functions. In *Proceedings of the 2002 World Congress on Computational Intelligence, FUZZ-IEEE'02*, pages 203–208. IEEE Press, 2002.