LACS system analysis on retrieval models for the MediaEval 2014 Search and Hyperlinking Task

Justin Chiu
Language Technologies Institute
School of Computer Science
Carnegie Mellon University, Pittsburgh, USA
Jchiu1@andrew.cmu.edu

Alexander Rudnicky
Language Technologies Institute
School of Computer Science
Carnegie Mellon University, Pittsburgh, USA
Alex.Rudnicky@cs.cmu.edu

ABSTRACT
We describe the LACS submission to the Search sub-task of the Search and Hyperlinking Task at MediaEval 2014. Our experiments investigate how different retrieval models interact with word stemming and stopword removal. On the development data, we segment the subtitle and Automatic Speech Recognition (ASR) transcripts into fixed length time units, and examine the effect of different retrieval models. We find that stemming provides consistent improvement; stopword removal is more sensitive to the retrieval models on the subtitles. These manipulations do not contribute to stable improvement on the ASR transcripts. Our experiments on test data focus on the subtitle. The gap in performance for different retrieval models is much less compared to the development data. We achieved 0.477 MAP on the test data.

1. INTRODUCTION
The amount and variety of multimedia data available online is rapidly increasing. As a result, the techniques for identifying content relevant to a query need to improve, to effectively process large multimedia data collections. There are existing works utilizing multi-modality for multimedia retrieval [7]; the ASR transcript is part of the multi-modality, which is similar to the Speech Retrieval framework. However, we believe there is more to be discovered on the Speech Retrieval part, especially the interaction between retrieval models and ASR transcripts quality. Established retrieval models are commonly used for the text retrieval. Applying the retrieval model to ASR transcripts is a standard approach for Speech Retrieval. However, there are fundamental differences between text documents and spoken documents, and different retrieval model might have different characteristics that can be beneficial, or harmful, for retrieval performance. Specifically, we examine word stemming and stopword removal, techniques that have been shown to be helpful in text retrieval. Can these techniques also help in speech retrieval? This question is the basis for our experiments[1]. We carried out two different sets of experiments on the development data to examine the difference between subtitle and ASR transcript. Each set of experiments investigates the effectiveness of different retrieval models and processing techniques. Due to the time constraint, we only submitted experiments on subtitle test data. We find that the performance gap observed on development data does not show up in the test data.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

2. EXPERIMENTAL SETUP
The MediaEval 2014 Search and Hyperlinking task [4] uses television broadcast data provided by the BBC, together with subtitles. We also tested on the ASR transcription provided by LIMSI [6] as a comparison to investigate how retrieval models and techniques interact with a different type of data. In all of the following experiments, the transcription is first segmented into smaller units with fixed length (60 seconds) according to the methods presented in [5]. We tested a stopword list from Indri toolkit that contains 418 common English words. We also used the Krovetz word stemming algorithm [8]. Finally, we tested three different retrieval algorithms: Unigram language-modeling algorithm (LM) [3], Okapi [9] retrieval algorithm (Okapi) and a dot-product function using TF-IDF weighting (TF-IDF) [10].

3. EXPERIMENTS ON DEV DATA
We first present our results on the development (dev) data, reporting the Mean Reciprocal Rank (MRR). The dev experiment is known item retrieval. The parameter for Okapi retrieval models is k1= 1.2, b= 0.75 and k3 = 7, and the μ for LM is 2500.

Table 1. MRR on subtitles for dev data

	LM	Okapi	TF-IDF
Baseline	0.265	0.279	0.296
Stopword	0.278	0.285	0.300
Stemming	0.295	0.344	0.355
Both	0.310	0.341	0.368

Table 2. MRR on ASR transcript (LIMSI) for dev data

	LM	Okapi	TF-IDF
Baseline	0.187	0.180	0.173
Stopword	0.167	0.175	0.160
Stemming	0.158	0.162	0.183
Both	0.157	0.177	0.183

From Table 1 and 2, we can observe the interaction between different processing and retrieval models. Stemming and stopword removal provides persistent improvement on subtitles. On the other hand, for the ASR transcript, these appear unstable. Aside from the difference due to recognition errors, one possible factor contributing to this phenomenon is the size of vocabulary. The vocabulary size for the subtitle is 251506, while the vocabulary

size for the ASR transcription is 83094, one-third of subtitle vocabulary. The lack of vocabulary, combining with stemming or stopword removal, can potentially decimate words in the transcript, hence harm the retrieval result. Another phenomenon we observed are a significant performance gap between different retrieval models. TF-IDF retrieval model outperforms LM and Okapi retrieval models, which was unexpected. Since the dev data is a known item retrieval task (For each query, there is only 1 matching speech segment), we suspect that dev data might have some bias in favor of the TF-IDF retrieval model. Another possible factor for superior performance on the TF-IDF retrieval model is smoothing. Both LM and Okapi retrieval model relies on smoothing parameters, but there is no smoothing for the TF-IDF retrieval model. If the data have a good number of exact matching between query and documents, TF-IDF may outperform other retrieval models due to the absence of smoothing.

4. EXPERIMENTS ON TEST DATA

The experiments on test data are ad-hoc retrieval task, which is no longer restricted to one result per query. Due to the time constraint, we only submitted systems based on subtitle data. Our submissions use both word stemming and stopword removal, as this setup gave the most promising result on the dev data. Results on test data are in Table 3.

Table 3. Result on test data

	LM	Okapi	TF-IDF
MAP	0.470	0.473	0.477
P@5	0.767	0.720	0.747
P@10	0.677	0.683	0.673
P@20	0.560	0.575	0.578

The performance gap between retrieval models is much smaller compare to dev data. Yet the trend is still the same: TF-IDF gives the best performance compared to other retrieval models. We suspect that the absence of smoothing contributes, and can explain the superior performance on TF-IDF. In a regular retrieval task, TF-IDF is not expected to outperform Okapi and LM consistently.

While processing the experiments on the test data, we noticed a difference between dev and test queries. The number of words in dev queries was greater than for test queries. Originally we thought that this might be a factor affecting performance on different retrieval models, but it does not appear to be an issue. Still, we suggest the characteristics of queries in dev and test data should be more consistent, so that the datasets are better matched.

5. ANALYSIS

We find that the TF-IDF retrieval model is the best of the three models tested. We believe this is because it does not do smoothing. However, generally speaking, smoothing can provide significant improvement on the standard retrieval task. We conducted experiments with LM retrieval model without smoothing; the resulting MAP on dev data is less than 0.05. So we can only assume that TF-IDF retrieval model could possibly find the correct way for processing the absent query word on our data. The possible reason for the performance gap on dev data is query text length. TF-IDF (which relies on exact word matching) is stronger than the other approaches. The test data has much shorter query length, so the gap is not as great as we observed on the dev data.

Research in the Spoken Term Detection community suggests using context for improving retrieval performance [1] or using retrieval system fusion [2]. We did not complete our experiments on using context for improving retrieval performance, but we tried system fusion approaches with our 3 retrieval models. The resulting system usually has the performance that's between the 2 fused systems. We conjecture that the three retrieval models we used in this work is in generally similar with each other, and fusion is not helpful due to lack of complementary.

6. CONCLUSION

We examined how different retrieval models interact with different text processing techniques such as word stemming and stopword removal, on subtitles and ASR transcript, two different forms on the dev data. We find that stemming and stopword removal can provide persistent improvement on the subtitle data, yet for the ASR transcript, these processing mostly harm performance except for stemming on the TF-IDF retrieval model. The result on test data shows that the difference on retrieval methods is not that significant when the retrieval task contains more possible targets. TF-IDF still has the best performance, which we believe is due to the absence of smoothing technique.

7. REFERENCES

[1] J. Chiu, and A. Rudnicky. Using Conversational Word Burst in Spoken Term Detection. In *Proc. of Interspeech 2013*. Lyon, France, 2013

[2] J. Chiu, Y. Wang, J. Trmal, D. Povey, G. Chen, and A. Rudnicky. Combination of FST and CN Search in Spoken Term Detection. In *Proc. of Interspeech 2014*. Singapore, 2014

[3] T. M. Conver, and J. A. Thomas. *Elements of Information Theory*. 1991

[4] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *Proc. of the MediaEval 2014 Multimedia Benchmark Workshop*. Barcelona, Spain 2014.

[5] M. Eskevich, and G. J. F. Jones. Time-based Segmentation and Use of Jump-in Points in DCU Search Runs at the Search and Hyperlinking Task at MediaEval 2013. In *Proc of the MediaEval 2013 Multimedia Benchmark Workshop*. Barcelona, Spain, 2013.

[6] J. L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication 37*, page 89-108, 2002.

[7] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-Example Event Search using MultiModal Pseudo Relevance Feedback. In *Proc. of International Conference on Multimedia Retrieval*, page 297. ACM, Glasgow, UK, 2014.

[8] R. Krovetz. Viewing morphology as an inference process. In *Proc. of SIGIR'93*, page 191-202, Pittsburgh, USA, 1993

[9] S. Walker, S.E. Robertson, M. Boughamen, G. J. F. Jones, and K. Sparck-Jones. Okapi at TREC-6: Automatic adhoc, VLC, routing, filtering and QSDR. In *Proc. of Text Retrieval Conference (TREC-6)*, pages 125-136, 1998

[10] C. Zhai. Notes on Lemur TFIDF model http://www.cs.cmu.edu/~lemur/tfidf.ps